

True lies

Attempts to fool the public into mistakenly believing that lie detectors work do not make for either good law enforcement or sound public policy.

If you are ever unfortunate enough to take a polygraph lie-detector test, you will be told by its administrator to answer a question falsely. For instance, you might be told to say 'no' when asked if it is Thursday, even though it is. This is to establish a baseline for lying, the examiner will say. Even though you may feel very little guilt about doing just as you are told, the sensors strapped to your body will register changes in pulse, breathing and sweating. Or so you will be informed.

In fact, the polygrapher does not even need to read the traces, because they are unlikely to show anything, according to polygraph training manuals. The purpose of this 'test' is to establish in your mind the infallibility of the machine — if you lie, it will catch you.

As a result, a polygraph exam is only as useful as the examinee thinks it is. A nervous subject is more likely to betray a lie as much through intonation or behavioural changes as through any physiological parameters. Polygraphers are trained to pick up on such cues, making the polygraph a useful tool for interrogation.

But that's not the same as calling it a lie detector. In fact, the instrument's reliability in detecting lies has been thoroughly discredited, most recently after an exhaustive review by the US National Academy of Sciences. In a 2003 report, an academy panel concluded that the tool was worse than useless for catching spies — its main purported national-security function. Spies can learn to defeat the machine, the

panel pointed out, and their passing score merely serves to bolster their credibility.

The Department of Energy responded by reducing the number of employees at US nuclear weapons laboratories who are required to take the exam. But other US government agencies have been slow to follow suit, arguing that the polygraph remains useful to them.

New detection technologies are being rolled out in response to current security concerns in the United States (see page 692). These include devices that monitor brainwaves and cameras that read heat signatures in the face. But there are alarming signs that the inventors of these tools are more interested in getting them into the field than in doing the research needed to see if they work.

As with the polygraph, the problem lies with the seductive appeal of technology. In the courtroom, juries are far less likely to question the results of a 'scientific' test than the testimony of a witness. Consequently, the polygraph is now banned as evidence in most states.

If the purpose of these tests is really intimidation, then any technology will do, as long as subjects believe that it works. But if the goal is to discover deceit, then the technology must be validated by solid research. If this is to be done, the US government will have to engage independent scientists — rather than people who work for the companies marketing the machines, or the agencies that plan to deploy them, as is currently the case. Only then will the truth come out. ■

Why China needs an NIH

Chinese biomedical scientists are right to push for a research agency that will distribute grants on the basis of peer review.

The emerging strength and influence of Chinese biologists, working at home and abroad, is not yet adequately reflected in the structure of research institutions in China itself. Last April, Zhu Chen, vice-president of the Chinese Academy of Sciences, who has been tipped to become China's next science minister, highlighted the issue and proposed the creation of a new research agency akin to the US National Institutes of Health (NIH).

Chen's proposal was quickly backed in a petition by 22 prominent researchers, most but not all of them Chinese, who implored the Chinese government to take it up (see *Nature* 425, 333; 2003). But the idea is currently bogged down in disagreements over whether it should embrace an intramural component — a set of government laboratories engaged in health research — or confine itself to distributing extramural grants.

In the United States, the NIH has distinguished itself over the decades by combining these two functions. The National Science Foundation, the main US non-biomedical science agency, has, in contrast, successfully confined its activities to the support of extramural work in the universities. There is an argument that the NIH's intramural labs have made a significant scientific impact in their own right, as well as helping its component institutes to stay abreast of their respective areas of interest.

China already has many laboratories, however, including those run by the Chinese Academy of Sciences, that receive support more-

or-less directly from the government. Its most pressing need is for a better extramural grant system, to fund research proposals to university scientists on a competitive basis. There is a lot to be said for involving specialists outside China — including the growing cadre of excellent, Chinese-born scientists working in the United States and Europe — in the review component of such a system.

Xiao-Fan Wang, a Chinese-born cancer researcher at Duke University in North Carolina, outlined these ideas at a symposium on 30 March on biomedical alliances between China and California, co-sponsored by *Nature* and the University of California, San Diego. Chen would have made the presentation himself, but he was denied entry to the United States after he was unable to get a visa.

The current system for funding the life sciences in China is plagued with inefficiencies and shortcomings. Some of these were highlighted by the country's slow response to the emergence of severe acute respiratory syndrome (SARS) in 2002, which encouraged Chen to propose a new biomedical research agency in the first place. The proposal has rapidly gained the support of researchers inside and outside China whose participation in peer review would make it work. If properly implemented, with minimal bureaucracy, full transparency and review of grant proposals by independent scientists, such an agency could greatly accelerate the development of biomedical research in China. ■





Drug doubts
Questions raised
over antidepressants
for children
p682



Search wars
Competition heats
up to be the Google
of science
p683



Arctic find
Siberian lake set
to give unique
climate records
p684



Packing problem
Proof of Kepler's
conjecture proves a
headache to review
p686

Push to protect whales leaves seafloor research high and dry

Rex Dalton, San Diego

A prestigious US research ship's schedule is in disarray after geophysicists were forced to abandon two recent projects because of concerns that they would harm marine mammals.

The *Maurice Ewing* — a 2,000-tonne vessel operated by Lamont-Doherty Earth Observatory at Columbia University, New York state — has been docked in Mobile, Alabama, for the past two months after the cruises were blocked.

Conservation officials turned the *Ewing* away from Bermuda last November and from Mexico in February. A cruise off Venezuela, originally planned for January, has also been delayed for three months, to avoid the seasonal migration of humpback whales there. The Bermudan and Mexican cruises had been planned for several years, involved several institutions and dozens of scientists, and their cost was expected to total about \$2.6 million.

Researchers on the *Ewing* use air-guns that discharge a sharp burst of compressed air into the ocean to generate sound waves. Geophysicists monitor the reflections from these waves to map rock formations deep under the ocean floor. Marine biologists are becoming increasingly concerned that the use of these devices could harm marine mammals.

In 2000, for example, three beaked whales were found stranded at the Galapagos Islands within 500 km of a *Ewing* cruise, and in 2002, two more were discovered during a *Ewing* project off the west coast of Mexico. Some marine biologists have argued that the strandings are connected with the noise from the devices (see *Nature* **425**, 549; 2003).

The 30-day Bermudan cruise was to study the volcanic rock that supports the islands, and the 40-day cruise off Mexico's Yucatan Peninsula would have looked at the Chicxulub crater, thought to have been formed by the impact of a meteorite 65 million years ago.



All at sea: whales demolish the *Maurice Ewing*'s cruise schedule.

A senior official from the National Science Foundation, which oversees the *Maurice Ewing* and its work programme, described the cancellations as an "embarrassment", attributable to a lack of proper preparation. "We were not on top of this," says James Yoder, director of ocean sciences at the agency. "We need to review everything we are doing before we leave the dock."

But officials at Lamont-Doherty deny lax planning. "It is unthinkable that we would engage in activities damaging to global ecology," says Michael Purdy, a geophysicist and the institute's director.

Bermuda's environment ministry says that it did not receive an application to approve the cruise until about a month before the ship arrived off the islands on 9 November. Bermudan officials say that applications should be made a year in advance. "The ship showed up without permission," says Jack

Ward, director of Bermuda's conservation services. "They didn't give us anything to work with. It is crazy. It was easy to say no." Bermuda's 200-mile territorial ocean zone had been declared a whale sanctuary in July 2000. "They said they wouldn't have come if they'd known that," says Ward.

The researchers debated whether to work from outside Bermuda's territorial waters, before giving up and heading back to port. The trip took a week, with operating costs of \$25,000 a day. "It was a huge waste of time and money," admits Marcia McNutt, director of the Monterey Bay Aquarium Research Institute in California, and the Bermuda cruise's chief scientist.

But events were repeated on 26 February, when the *Ewing* arrived at Progreso, Mexico, to study the meteorite crater. Officials at the Secretariat of Environmental and Natural Resources, an environmental agency of the Mexican government, were unconvinced that the cruise would be safe for marine mammals, and ordered the ship to leave Mexican waters. It returned to dock in Mobile.

Sean Gulick, a geophysicist at the University of Texas at Austin and chief scientist on the Mexico cruise, was unavailable for comment. But Paul Stoffa, director of the university's Institute for Geophysics, says that future cruises will have to prepare more carefully. "The mode of operation should be to go through all the steps well in advance," he says.

Another team, which was planning to take the *Ewing* to Queen Charlotte Sound near Vancouver, Canada, in October, will delay the trip until October 2005 to ensure it has the support and approval of local agencies and populations, says project leader Lincoln Hollister, a geologist at Princeton University.

The actual impact of the experiments on marine life remains uncertain, however. The US Marine Mammal Commission, which is responsible for monitoring marine mammal issues, has formed an advisory committee to assess the problem.

French government concedes defeat to researchers

Declan Butler, Paris

Scientists gathered outside the Sorbonne last week, not to demonstrate, but to picnic and toast with champagne a victory in their three-month conflict with the government.

The Save Research movement, which had led a series of protests (see *Nature* 428, 241; 2004), called them off on 7 April after the government caved in to the researchers' demands.

The move came after French president Jacques Chirac's statement on 1 April, when he disowned the research policies of his previous government and declared the scientists' protests "justified". François Fillon, minister for education, higher education and research, and François d'Aubert, junior minister for research, who were appointed in a government reshuffle on 31 March,

moved quickly to put Chirac's words into action.

Following negotiations last week with representatives of the scientific community, Fillon and d'Aubert announced a series of emergency measures for research,

including the scientists' key demand of 550 new full-time research posts for young scientists. The government also agreed to a further 1,050 university posts — 300 immediately and 750 in January 2005.

Fillon described the move as "an exceptional and immediate effort in favour of science jobs", and a means "to take the high ground" in defusing the situation. Finding a way out of the current crisis was imperative, he added, if progress is to be made on a planned reform and funding plan for French research that is expected to go before parliament by the end of the year. He also committed to the protesters' demands that the scientific community be brought on board in drafting the reforms.

"It's a great day for French science," said a jubilant Alain Trautmann, leader of Save Research. "We got exactly what we demanded." ■



Alain Trautmann celebrates Save Research's triumph.



Depressed youngsters may get more benefit from behavioural therapy than from antidepressants.

Trial analysis questions use of antidepressants in children

Erika Check, Washington

Children taking antidepressants are unlikely to reap much benefit, according to researchers in Australia. They say instead that data from clinical trials, including some whose results helped win approval for the drugs, suggest that the treatment offers few advantages.

In a review published in the *British Medical Journal* on 10 April, child psychiatrist Jon Jureidini of the University of Adelaide, Australia, and his colleagues reanalysed the data from five trials of popular antidepressants, known as selective serotonin reuptake inhibitors, or SSRIs (J. N. Jureidini *et al.* *Br. Med. J.* 328, 879–883; 2004).

"We found that a superficial look at the papers gives you one impression, but a deeper look gives a different impression," Jureidini says. "Very little benefit and the possibility of significant risks should equate to a good deal more caution in the use of the drugs."

The review appears as regulators and patient advocates around the world are taking a closer look at the drugs' safety and efficacy in children. British officials have cautioned doctors not to prescribe them to young people, and US regulators have strengthened warnings about the drugs' potential to increase the risk of suicide.

Jureidini's group looked at five published trials of paroxetine, fluoxetine and sertraline in children. These drugs are sold as Paxil, Prozac and Zoloft, respectively, in the United States, and as Seroxat, Prozac and Lustral in Britain, and were prescribed more than 6 million times to US children in 2002 alone.

Two of the studies — one of fluoxetine, one of sertraline — had reported that the drugs helped children to overcome depression. But when Jureidini and his colleagues

grouped the data on the five trials together, they found that the drugs as a class had only a marginal effect — the equivalent of an improvement of 3–4 points on a depression scale that ranges from 17 to 113.

They also found that children in the placebo groups of the studies improved almost as much as children in the treatment groups. And significantly more children taking paroxetine and sertraline dropped out of the trials because of adverse events than those in the placebo groups, they report.

Jureidini and his colleagues also contend that the drugs' popularity is diverting patients from treatments that could help them more, such as cognitive behavioural therapy, which teaches patients about their thinking patterns and their reactions to difficult situations.

Authors of the published studies say that grouping the data from positive and negative studies together dilutes the positive findings of individual studies.

"If you pool all these studies and average them out, you get less significance," concedes Graham Emslie, a psychiatrist at the University of Texas Southwestern Medical Center at Dallas, who led two studies of fluoxetine that were scrutinized in Jureidini's review. But Emslie says that the antidepressants still work better than any other treatment studied. "It concerns me that they are suggesting that people put more faith into unproven treatments than into treatments that have actually been studied," he adds.

Emslie also says that the only treatment that might work better than antidepressants is cognitive behavioural therapy, and that he is leading an ongoing study, due to be completed in about three months, that compares the benefits of this with those of medication. ■

Publishers go head-to-head over search tool

Jim Giles, London

Will there ever be a science equivalent of Google? Two of the world's biggest science publishing and information firms seem to think that there will. They are about to compete head-to-head to create the most popular tool for searching the scientific literature.

Elsevier, the Amsterdam-based publisher of more than 1,800 journals, has announced that this autumn it will launch Scopus, an online search engine covering abstracts and references from 14,000 scientific journals. Scopus will arrive as a direct competitor for the established Web of Science, owned by Thomson ISI of Philadelphia, the scientific information specialist.

"Scopus will definitely be a threat to ISI," says one science publishing expert, who asked not to be named. "But ISI will not just let this happen. There will be some kind of arms race in terms of adding new features."

Many researchers are already wedded to subject-specific databases of scientific information, such as PubMed, for biomedical research. But Web of Science is currently the only service to cover the full spectrum of scientific disciplines and publications. It can also generate the citation statistics that are sometimes used to measure the quality of journals and individual papers.

ISI, which is widely used by libraries worldwide, may be hard to displace. It covers fewer than 9,000 journals, but it has been available in its present form since 1997 and includes a 60-year archive of papers. Thomson ISI says it will extend this to 105 years by



Definitive database? Scopus and Web of Science will help scientists search the literature.

the end of 2005. The company also owns the only extensive database on patent abstracts.

Elsevier cannot hope to match this coverage in the short term. The company has been able to draw on its experience of running biomedical and pharmaceutical databases, and developers began compiling a multidisciplinary

index two years ago. Even so, when it launches, Scopus will index only five years of references for some journals, rising to ten years during 2005. Data on abstracts will go back further, in some cases to the mid-1960s.

Because Scopus has been built from scratch, Elsevier has been able to work with librarians to develop an alternative to the Web of Science interface, which has been criticized by some users. "Users are very happy with Scopus," says Steven Gheyselinck, a librarian at the University of Lausanne in Switzerland who has been testing it.

Although Scopus and Web of Science are the only products aiming to cover all of science, other search engines are also under development. The Google of science could end up being Google itself: the company has collaborated with nine publishers, including Nature Publishing Group, to create an engine called CrossRef Search.

This service, a pilot of which appeared last month, allows users to search digital versions of all papers held by the publishers involved and returns links to articles on their websites. Unlike Web of Science and Scopus, which scan through the titles and abstracts of articles, CrossRef Search also searches the full text of papers. Many of the other 300 or so members of CrossRef — a publishers' collaboration established to allow easier linking between citations — are likely to join the service if the pilot is successful. ■

♦ www.isinet.com

♦ www.crossref.org/crossrefsearch.html

♦ www.scopus.com

Six-day sacking over as researcher regains Italian job

Alison Abbott, Munich

It was no joke when Lucio Luzzatto was told on 1 April that he was being sacked as scientific director of Italy's National Cancer Institute in Genoa — even though he was reinstated by the health ministry just six days later.

The institute's temporary commissioner, Maurizio Mauri, sacked Luzzatto after accusing him of breaking the terms of his contract by acting as a consultant for his former employer, the Memorial Sloan-Kettering Cancer Center in New York. Mauri later accepted that only industrial consultancies were disallowed.

Luzzatto says that he was the victim of a battle over whether a research institute should be run by scientists or administrators. But many of his supporters say he fell foul of the Byzantine politics that permeate Italian science and academia.

Mauri was appointed in 2001 to hold the

reins while a new management structure was implemented at the National Cancer Institute, one of Italy's 32 clinical research institutes, whose restructuring has been delayed for more than a decade. During this time, each institute has been in the hands of commissioners who have full administrative control. A law was passed last year to let the institutes run themselves again but this has yet to be implemented.

Mauri charged that Luzzatto broke the terms of his contract by working with the Sloan-Kettering centre, which he left in 2000 to take up his post in Genoa. Luzzatto argues that such collaborations are both normal and desirable in academic research, and points out that he spent a total of only 16 days in New York last year. Scientists throughout Italy wrote to the health ministry in his defence, attesting to his dedication to the Genoa institute.

At a meeting on 6 April, the ministry

accepted Luzzatto's arguments and Mauri retracted his letter of dismissal. But Luzzatto says he remains under strong pressure to leave the institute voluntarily. In an attempt to defuse the tensions, the local government of Liguria has offered to finance an independent laboratory for him.

Luzzatto suspects that the real reasons for the dismissal were his disagreements with Mauri about how the institute should be run. One such dispute concerned Mauri's decision to claw back salary supplements from PhD scientists that are legally applicable only to those with MDs. Luzzatto had argued that it would be demoralizing to have different pay for the same work within a lab.

"Italy says it wants to reverse the brain drain — it is appalling that a top scientist who returns to Italy should be treated in this way," comments Giovanni Romeo, a medical geneticist at the University of Bologna. ■

Arctic lake promises hot data on past climate

Quirin Schiermeier, Munich

A preliminary excursion to a remote Russian lake has raised scientists' hopes that it will offer the most reliable terrestrial records yet of arctic climate history.

Located in the extreme northeastern tip of Siberia, Lake El'gygytyn is unique, climate researchers say. Unlike most of the Arctic, the lake has never been covered by glaciers, which disrupt the accumulation of sediment. This means that El'gygytyn's floor offers an uninterrupted view of past climate patterns.

"Drilling cores from the lake should provide a fantastic climate archive," says Thomas Stocker, a climate researcher at the University of Bern in Switzerland. "It would be a valuable supplement to marine drillings, whose information is always filtered by the ocean."

During a four-month expedition last year, geologists and limnologists from Germany, Russia and the United States probed the lake's sediments and surrounding permafrost soil to check out their likely usefulness. And at a meeting in Leipzig, Germany, late last month, they announced that their study had confirmed the lake's potential.

"Our survey has revealed that the bedding of the lake floor sediments is perfectly undisturbed," says Marin Melles, a geologist at the University of Leipzig and member of the expedition.

The lake is now being put forward as a site for the International Continental Drilling Program, a collaboration of 11 countries that aims to extract cores from the continental shelf, which would conduct a 400-metre deep-drilling project there in 2007.



Icy yield: drilling at Lake El'gygytyn has already retrieved a climate record for the past 400,000 years.

El'gygytyn, which is about 175 metres deep and 15 kilometres across, fills the larger part of a crater created by a meteorite impact 3.6 million years ago. It is on the Chukotka peninsula, opposite Alaska, an area far from any settlement and very difficult to reach.

The expedition team and nine tonnes of equipment were flown to the lake last April by helicopter from Pevek, a small town on the East Siberian Sea.

Getting to the lake was difficult, Melles says. At first, Russian customs in St Petersburg declined to clear several items, including a snowmobile, which never made its way east. And summer evacuation from the camp, which was beginning to drown in seasonal mud, was delayed for a few days because available Russian helicopter pilots had hit statutory limits on their monthly flying hours. To get out, the researchers eventually had to hire a helicopter from a local gold-mining company.

But the scientific yield was worth the effort, says Julie Brigham-Grette, a geologist at the University of Massachusetts in Amherst. Brigham-Grette, in collaboration with Melles and with researchers at the North-East Interdisciplinary Scientific Research Institute in Magadan in far-eastern Russia, is analysing mineralogy and organic traces in the lake-floor sediments.

Seismic profiling shows that the sediments are some 400 metres thick, and probably contain an exceptional record running from the meteorite impact, in the Pliocene warm period, up to the present day. The expedition recovered a drilling core 16 metres long, representing the past 400,000 years, from the centre of the lake.

Deep drilling in Lake El'gygytyn would help researchers study arctic climate change before the onset of glaciation 2.2 million years ago, explains Brigham-Grette.

♦ www.icdp-online.de/welcome.html

Queen flies the flag for cancer alliance at Paris bash

Declan Butler, Paris

Queen Elizabeth II was in Paris last week to mark a century of Anglo-French accord. And as part of the celebrations she found herself rubbing shoulders with top cancer researchers from both sides of the Channel.

A sumptuous dinner at the British Embassy in Paris saw the Queen, her husband Prince Philip and Bernadette Chirac, wife of French president Jacques Chirac, assuming the role of matchmakers to bring scientists from the two countries closer together. The event was held to celebrate the centenary of the *entente cordiale*, an agreement signed on 8 April 1904 to end the countries' colonial disputes.



All abroad: the Queen arrives in Paris.

On the menu, apart from a 1985 *Château Margaux*, was cooperation on cancer. British figureheads such as Colin Blakemore, head of the Medical Research Council, Mark Walport, chief executive of the Wellcome Trust, and Liam O'Toole, director of the National Cancer Research Institute, met the leaders of France's top cancer centres, such as Daniel Louvard, director of the Curie Institute, and Thomas Tursz, director of the Gustave Roussy Institute.

John Holmes, the British ambassador in Paris, announced a €600,000 (US\$730,000) scheme, funded equally by Britain and France, to support the best young cancer researchers from both countries. And the Princess Zahra

Aga Khan, daughter of businessman the Aga Khan and coordinator of the health services arm of the Aga Khan Development Network, agreed to fund an Entente Cordiale Cancer Research Prize to reward one up-and-coming researcher from each country.

Cancer leaders at the event said that both countries could benefit from collaboration in everything from bioinformatics to tissue banks. Their information is currently held in different languages, on incompatible computer systems.

Researchers also agreed to a joint conference this autumn on bioethics — an area where Britain's utilitarian outlook is sharply at odds with the more philosophical French position (see *Nature* 389, 661–662; 1997). The two countries will organize rock concerts, operas and sporting events this year to fund their cancer-research plans.

Researchers cheered by hunt for genetic basis of depression

Munich A project to probe the genetics of depression was launched last week at the Human Genome Meeting in Berlin, Germany. Named NEWMOOD, the project has €7.3 million (US\$8.9 million) from the European Union to fund research in 13 laboratories in ten countries over the next five years. It aims to aid the development of novel drugs against depression, using mouse and rat models.

“Antidepressant drugs haven’t changed much for the past 30 years,” says the project’s leader, Bill Deakin of the University of Manchester, UK. “We have to find new molecules that are involved in depression so that new treatments can be developed.”

Most drugs for depression boost levels of the brain chemical serotonin. But this can take weeks to take effect, and even then works for only around half of patients.

Journal juggles balls to publish Kepler proof

Washington After more than five years in peer review, a proof of a famous mathematical problem is finally being published.



Orange squash: greengrocers routinely pack fruit into the densest possible arrangement.

The solution to Kepler’s conjecture, which states that the densest possible arrangement of spheres is to stack them in a pyramid — like oranges in a shop display — was announced in 1998. But the proof, developed by Thomas Hales of the University of Pittsburgh, used computers to check every possible arrangement. Referees and editors at the prestigious *Annals of Mathematics* found it impossible to check all the code and its outputs, and were unsure how to verify the paper (see *Nature* 424, 12–13; 2003).

The editors have now decided on a compromise solution. The code will appear only on the journal’s website and a note will

be published explaining that it has been impossible to verify in full. The rest of the proof — the logic behind the writing of the code — will appear as a normal paper in the print edition.

Climate change blamed for making sea sick

San Francisco The seas are getting sicker, warn researchers who have studied reports of marine disease from the past 30 years.

Disease rates have risen in five of nine marine groups analysed — turtles, corals, mammals, urchins and molluscs (J. R. Ward and K. D. Lafferty *PLoS Biol.* 2, 542–547; 2004). Study author Jessica Ward, who researches marine disease at Cornell University in Ithaca, New York, suggests that climate change and the introduction of terrestrial pathogens, such as the canine distemper infecting Antarctic seals, could be to blame.

Scientists already knew that there have been more papers on marine disease in recent decades, say Ward and co-author Kevin Lafferty of the University of California, Santa Barbara. But it was not known if this just reflected increasing interest in the subject. By normalizing these records for increases in research activity, Ward and Lafferty concluded that the rise is genuine.

Endangered Tahitian snails found in Michigan freezer

Sydney The chance discovery of a batch of frozen Tahitian land snails in a freezer at the University of Michigan could help researchers study and conserve the almost-extinct animals.

The samples were collected in the 1970s for a study of the molluscs' evolutionary diversification. But they were forgotten

when one of the project's members died. The species have since been driven to near-extinction by the rosy wolfsnail (*Euglandina rosea*), which was introduced to Tahiti to control pests but ate the native snails instead. It was a "spectacularly inept attempt at biological control", says Diarmaid Ó Foighil, a mollusc expert at the University of Michigan, Ann Arbor.

Ó Foighil is extracting DNA from the rediscovered samples to reconstruct the snails' family tree. He hopes that the results will aid their conservation.

Animal database plays tag with DNA barcodes

San Diego A DNA-barcode database of animals is to be created by an international consortium of museums to support research projects.

The Barcode of Life Initiative is initially being funded to the tune of \$670,000 by the Alfred P. Sloan Foundation of New York, to aid the identification and documentation of animals in projects involving environmental genomics, biomedicine and endangered species.

A secretariat is being established at the National Museum of Natural History in Washington DC, which will host a meeting in May on ways to link the

database to museum specimens.

The team plans to categorize animals using a fragment of a mitochondrial gene, cytochrome c oxidase subunit I. In the future, a plant database is to be created based on a different gene.

UK space scientists gamble on lottery for funding

London Britain's tight space-science budgets have prompted scientists to turn to gambling in a bid to boost the country's research effort.

Money from the Aurora Lottery Syndicate, which costs £10 (US\$18) to join, will be used to buy tickets for the weekly draw of the UK national lottery. Any winnings above £500,000 will be donated to the government to fund Britain's participation in the Aurora Programme, a long-term European Space Agency project for Solar System exploration. Groups promoting British space science will also receive part of any big wins.

Syndicate organizer Julia Tizard, a cosmochemistry PhD student at the University of Manchester, says she hopes to have 100 members after the scheme is launched on 16 April. The odds of matching the six winning lottery numbers are 1 in 13,983,816.



Cool customer: rare Tahitian land snails have been lost in a freezer since the 1970s.

J. B. BURCH

G. LINDSAY



Slaves to the bench: they may be the backbone of US research, but many postdocs endure poor employment conditions.

Young, gifted . . . and broke

At the turn of the millennium, the US National Academies put the spotlight on the miserable pay and conditions experienced by most US postdocs. Things are now starting to change, but slowly. Betsy Mason reports.

Low pay, a lack of benefits, inadequate recognition and poor career guidance. This, according to an influential report¹ released in 2000 by the US National Academies, is the lot of postdocs working in the United States. Across the country, more than 50,000 scientists and engineers make up this part of the workforce, existing at that indefinite stage of continued training between earning a PhD and gaining a permanent academic position.

These young researchers represent the engine room of US science, responsible for most of the hands-on work that underpins papers published each week in leading journals. And if the most talented of them are forced to quit academia for greener pastures, they will take with them the vibrancy that drives US scientific enterprise.

This week in Washington DC, the academies' Committee on Science, Engineering, and Public Policy (COSEPUP), which produced the report, is holding a conference to review progress since it thrust the problems faced by postdocs into the limelight. In some respects, the news is good. A National Postdoctoral Association (NPA) has been set up to represent postdocs' interests. Local associa-

tions have sprung up at academic institutions across the country. And for some postdocs, salaries and benefits have improved.

But problems still abound. Some universities are conspicuously lagging behind in improving pay and conditions for their postdoctoral researchers. And for those postdocs — more than half of the total — who are visitors from abroad, visa restrictions imposed under the new focus on homeland security are causing serious difficulties².

"It's a mixed bag," says Timothy Coetzee, director of research training programmes at the National Multiple Sclerosis Society, a charity that has won plaudits for its enlight-

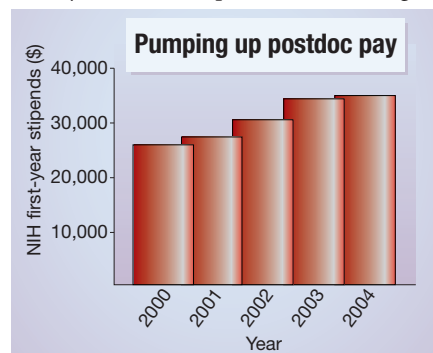
ened attitude towards career development for the postdocs supported by its grants. "We've made some progress, but we still have a long way to go."

On the rise

The National Institutes of Health (NIH), which funds more postdocs than any other agency or institution in the country, has also taken notice of the situation. "It's an issue of considerable importance to us," says Walter Schaffer, acting director of the agency's Office of Extramural Programs.

Since 2000, the NIH has increased the minimum stipend for first-year postdocs receiving its National Research Service Awards, of which some 1,800 are given out each year, by almost one-third (see Graph, left). Those funded under the scheme are also guaranteed a salary rise of about 5% with each year of experience. This is good news for other postdocs as well, because many institutions, including major research universities, are adopting the NIH pay scale³.

Another positive sign is the birth of the NPA, founded in January 2003 by a group of postdocs in the biological sciences looking for an organized voice at the national level.



SOURCE: NIH



Trailblazer: Stanford University was one of the first to set up a dedicated postdoc office.

The group tackles issues such as the need for better salaries and benefits, and is trying to gather data on current institutional postdoc policies. The NPA is already being taken seriously by agencies such as the NIH, which invited the group's representatives to take part in a meeting on postdoc training and opportunities in October last year.

"This is a good time to deal with postdoc issues because everybody is interested right now, and they're listening," says Claudina Stevenson, a founding member of the NPA who works at Harvard Medical School's Dana-Farber Cancer Institute in Boston, Massachusetts.

The NPA is also helping postdocs to form local associations, and encouraging institutions to create offices that specifically deal with policies concerning postdocs. It estimates there are now more than 50 institutions with postdoc associations or offices, most of which came into existence in the past few years.

Stanford University in California was one of the pioneers, creating a postdoc office in 1989. Ten years later, its postdocs also formed their own association. Stanford's 1,300-plus postdocs enjoy salaries that are higher than the NIH minimum and have benefits that include health insurance and maternity leave. Other success stories include the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts, where postdocs have dental insurance and get a supplement of \$2,000 per year to help them plan for their retirement. And at the University of North

Carolina at Chapel Hill, the local postdoc association worked with the postdoc office to establish a minimum salary and a five-year cap on the time a researcher can be employed as a postdoc — if they are kept on longer, they must be given a permanent position.

At the University of Connecticut, meanwhile, postdocs have formed a fully fledged labour union that bargains with the administration on salary and benefits. Last month, it negotiated a rise in the minimum postdoc salary from \$27,000 to \$34,200, and won annual rises, health insurance, paid sick leave and holidays, plus a standardized grievance procedure. If the Connecticut legislature doesn't object, these policies will become official next month.

But at many institutions, postdocs still toil without a guaranteed salary minimum. "Awareness of postdoc issues has definitely improved," says Helen McBride, a fifth-year postdoc in developmental biology at the California Institute of Technology in Pasadena, where the high cost of living means that the minimum salary of \$32,000 doesn't go far. "But there's still a lack of this awareness translating into change."

Second-class citizens

Recipients of postdoctoral fellowships are often classified as 'trainees' at their host institutions, which can leave them ineligible for benefits such as healthcare. For instance, Cornelia Bentley and her husband Jeffrey Pitman, now both postdocs in molecular biology at Brown University in Providence, Rhode Island, were paying \$680 per month to insure their family of three because Pitman's fellowship from the American Cancer Society left them without health coverage. When they couldn't make ends meet, the principal investigator of Pitman's lab stepped in temporarily to cover the insurance costs until Bentley managed to secure a postdoc with benefits. "I don't know what we would have done otherwise," Bentley says.

Such problems are becoming more acute because of a congested job market that leaves too many PhDs chasing too few permanent academic positions. As a result, many researchers have to eke out a living as postdocs into their late thirties or early forties, or just give up and look for opportunities elsewhere.

More and more postdocs are switching to careers in industry, and some are leaving science altogether. After spending six years as a postdoc in genetics at the Wadsworth Center in Albany, New York, and having no luck landing a faculty position, Derek Scholes is now applying for fellowships in bioethics. "I typify many postdocs who were keen to enter a research career and for whatever reason, didn't get enough publications to be competitive," he says.

Some researchers fear that the grim job outlook, combined with poor pay and conditions, may discourage the very best students

from entering science in the first place. William Zumeta, who studies education policy at the University of Washington in Seattle, has found some evidence that the best physical-sciences graduate students are opting to go to business school instead⁴. "If you buy the idea that the best and the brightest drive scientific innovation, this could be a real problem," he says.

A foreign affair

Until recently, US researchers could count on a boundless supply of high-quality foreign scientists to fill positions in their labs. But tighter security has made it tougher for such researchers, particularly those in fields deemed sensitive by the US government, to get visas. This topic is likely to be hotly debated at this week's COSEPUP meeting.

Another live issue is the tendency of some principal investigators to view their postdocs as little more than hired hands. "There is a lack of understanding of what a postdoc is," says one evolutionary biologist, who left her first postdoc after a year when she realized her adviser was treating her like a technician.

What's needed, say postdoc activists, are policies to ensure that postdocs receive proper training and career guidance designed to aid their transition into independent researchers supervising their own students and fellows.

Some funding bodies are now providing a small number of grants with this express aim. The National Multiple Sclerosis Society awards three 'career transition fellowships' each year, which are worth about \$550,000 and consist of two years of postdoctoral funding followed by salary and research support for the first three years in a faculty position. Another research foundation, the Burroughs Wellcome Fund, has awarded 15 similar career grants since 2002. But these schemes are a drop in the ocean — many more such grants will be needed to make a difference to the academic job market. "The big problem is that policy-makers don't want to take responsibility for career development," says Zumeta.

As this week's COSEPUP meeting assesses progress, the grades it marks on its report card are likely to range from A to F. Postdocs still have a long list of grievances, but at least things are moving in the right direction. "I think we've done most of the easy stuff," says Coetzee. "Now comes the hard part of deciding where to put our priorities." ■

Betsy Mason is a freelance writer based in New Haven, Connecticut.

1. *Enhancing the Postdoctoral Experience for Scientists and Engineers* (National Academies Press, Washington DC, 2000).
2. Brumfiel, G. *et al. Nature* 427, 190–195 (2004).
3. Kreeger, K. *Nature* 427, 178–179 (2004).
4. Zumeta, W. & Raveling, J. S. *The Best and Brightest for Science: Is There a Problem Here?* (Commission on Professionals in Science and Technology, Washington DC, 2002).

COSEPUP meeting

♦ www7.nationalacademies.org/postdoc

National Postdoctoral Association

♦ www.nationalpostdoc.org



The truth about lying

Crooks, terrorists and liars can be hard to spot. But some researchers hope that scanning brains, faces or voices might reveal deceivers. Jonathan Knight looks at lie-detection technology, and wonders who is being fooled.

When talk-show hosts and death-row inmates come knocking on his door, neuroscientist Lawrence Farwell welcomes both with open arms. Farwell is the inventor of a brain fingerprinting machine, and since he began consulting in high-profile murder cases four years ago, he has found no shortage of publicity for his technique. In March alone, Farwell featured in a Public Broadcasting special and appeared on ABC's *Good Morning America* and CNN. "They come to us," says his public-affairs manager. "All I do is sit here and field phone calls."

Farwell's true quest is for credibility, not publicity. Brain fingerprinting is far from being an accepted forensic tool, despite claims on Farwell's website that it was used to trap a serial killer into admitting his guilt and to help free a man jailed for 25 years for a murder he says he did not commit.

But while the media rush to book Farwell, some scientists say his claims are overblown. Farwell says brain fingerprinting can determine with near certainty whether specific information is stored in a person's brain. The critics say the technique, like most others used to sort truth from lies, suffers from a lack of peer-reviewed evidence that it works.

"The necessary research has never been done," says Emanuel Donchin, a psychophysiological at the University of South Florida in Tampa and Farwell's former graduate adviser.

Yet these unproven technologies are becoming increasingly attractive to US law-enforcement and security agencies. The traditional tool of lie detection, the polygraph, has come under withering criticism in the past decade. With growing national security concerns, the hunt for alternatives has intensified. Laboratory tools — from infrared sensors to eye trackers — are being converted into lie detectors to help catch criminals and spot terrorists before they strike.

Sweaty palms

"We are no longer just looking for specific acts of deception," says Andrew Ryan, director of research at the Department of Defense Polygraph Institute (DoDPI) in Fort Jackson, South Carolina. "We are interested in the overall credibility of people. It's a much broader field."

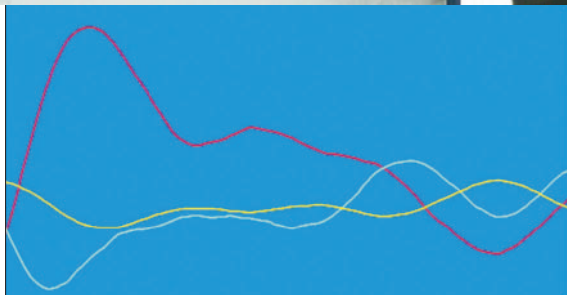
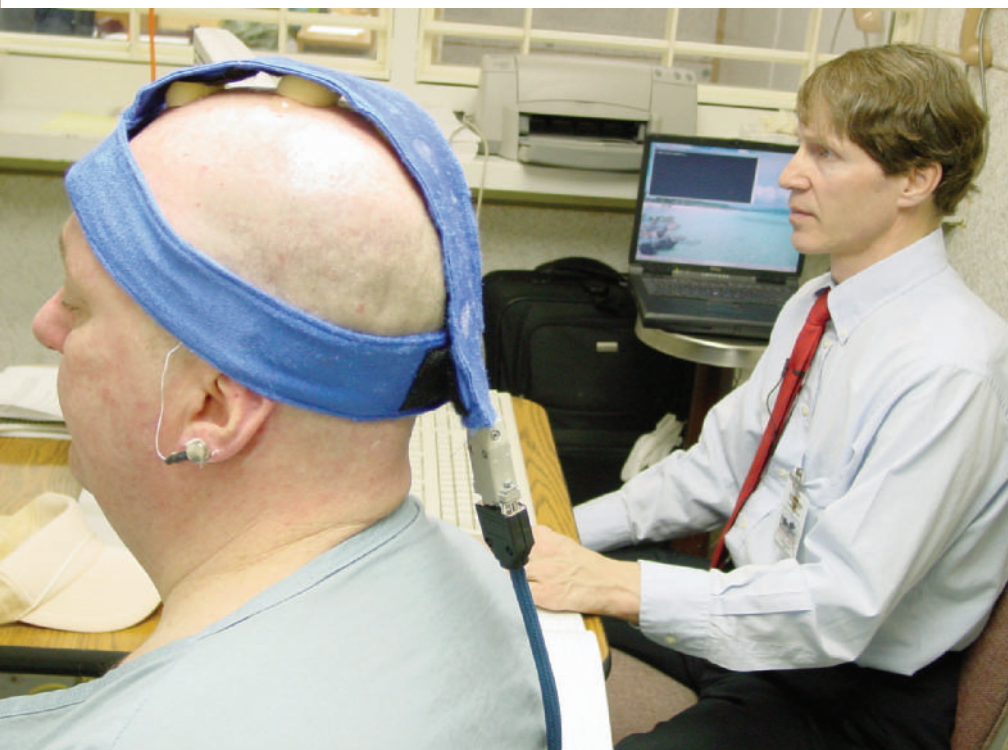
The credibility of the polygraph has long been questioned. Conceived in 1915 by psychologist William Marston, who also invented the comic-strip superhero Wonder

Woman, the polygraph measures changes in perspiration, breathing, pulse and blood pressure as the subject answers a series of yes-or-no questions. When we lie, the theory goes, changes in these parameters betray us.

Certainly the stress of lying in an interrogation can cause physiological changes, from sweaty palms to a rapid heart rate. But then so do other high-pressure situations, such as being suspected of a crime or being hooked to a machine and asked a lot of questions. As a damning 2003 report on polygraphy by the National Academy of Sciences concluded, there is little scientific evidence that polygraphers can identify dishonest responses with any certainty¹.

In response to the academy report, and the concerns of employees, in September 2003 the US nuclear weapons labs reduced the number of routine polygraph tests from 20,000 to 4,500. Other agencies still use the polygraph for security screening and in law enforcement, but rarely as evidence in trials. A 1998 US Supreme Court ruling upheld state restrictions on the use of the polygraph in courtroom proceedings, and today most states ban it altogether.

It is this gap that Farwell hopes brain fingerprinting will fill. The technique does not



Lies, damned lies, and technology: the accuracy of the polygraph lie detector (facing page) has been attacked. Lawrence Farwell (above, right) thinks measuring the burst of brain activity that occurs when a person recognizes something (left) could provide a reliable alternative. Others fear the technique works better in the lab than in the field.

measure deception, merely whether information is stored in the brain. Farwell's subjects wear standard electrodes for recording brain activity, and are shown a series of words on a computer screen, some of which refer to details of a crime that only the police and someone present at the scene would know. If the murder weapon was a gun, the suspect might see the words Luger, Remington and Uzi. If the suspect knows the murder weapon was a Remington, seeing the word should evoke a 'brainwave' that the electrodes can detect.

It sounds far-fetched, but the brainwaves Farwell relies on are well documented. Also known as event-related potentials or ERPs, they measure the electrical activity of myriad neurons in response to a stimulus, such as a word or an image. ERPs have been studied for decades, but the cognitive processes underlying them are still unclear.

One bump in the ERP trace, called a P300 because it occurs about 300 milliseconds after an event, generally occurs when the stimulus holds special significance for the subject. In 1991, Farwell and Donchin showed that in a controlled laboratory setting the P300 could be used to reveal hidden knowledge of a mock crime conducted just

prior to the test². Later, Farwell developed a related measure called a MERMER, which he went on to patent³. Four years ago, he set up Brain Fingerprinting Laboratories, a company now based in Seattle, to commercialize the technology for both civilian and national-security uses.

Brainwaves

But critics charge that the technology is not ready for use in the real world. Peter Rosenfeld, a psychophysicist who studies ERPs at Northwestern University in Evanston, Illinois, points out that in the only published field test of ERPs for detecting deception, the machine did little better than guessing⁴. There are many factors that could affect the outcome. Because the P300 is measuring recognition, whether you remember an event is crucial. For instance, how does memory of a crime change over time, or in response to drugs or

stress? Does a psychopath even have a normal P300 response? Such factors need to be properly tested, Rosenfeld says. "It will take a substantial amount of research before this is ready."

And what about false positives? Say the suspect is innocent, but collects Remingtons. It doesn't matter, claims Farwell, because they test memory of several crime-related items, not just one. William Iacono, a psychologist at the University of Minnesota in Minneapolis, agrees that it is hard to imagine how getting six out of six hits, for instance, could be a false positive: "How do you explain that just by chance?"

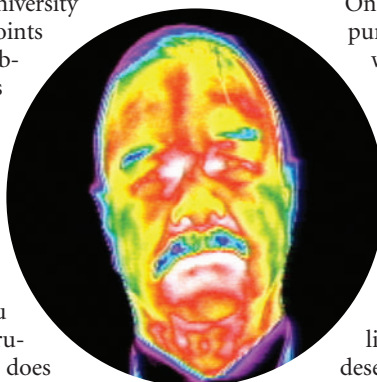
Farwell maintains that it is better to use the test and let a judge or jury decide how much weight to give it. "I don't agree that you have to remove every possibility of a false positive or a false negative," he says. "Court cases get decided every day on 60:40 probabilities. What we bring to the judge and jury is highly accurate information."

Hot flushes

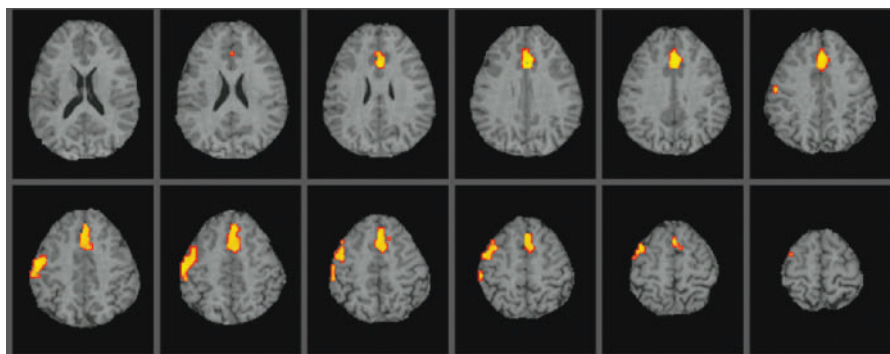
The P300 response is supposedly beyond the subject's control, but there may be ways to manipulate it. In a newly published study, supported by the DoDPI, Rosenfeld's group exposed several students to a mock crime and then gave them a P300 test. Some students had instructions to imagine their professor slapping them in the face during each stimulus. As a result, two-thirds of them were able to produce P300 waves even for stimuli that should have been negative⁵. This approach wouldn't clear a suspect who used it, but it would render the test useless.

The security agencies desperately need new tools to help them identify terrorists. Can brain fingerprinting help? Farwell says he has been able to tell FBI agents from civilians by checking for P300s elicited by terms learned during FBI training. So in theory, a P300 test could pick out a terrorist by finding records in the brain of something only a terrorist would know — phrases from an al-Qaeda training manual, for instance. The trick is finding information that no innocent person would know.

One approach the DoDPI is pursuing is an eye tracker, which, like brain fingerprinting, checks for concealed knowledge. The theory is that the eyes involuntarily spend less time scanning familiar objects than unfamiliar ones, a difference measurable in milliseconds. So the eyes of a terrorist will linger less over a photo of the desert camp where he trained than over a photo of a survivalist camp in Nebraska. A prototype



Trial by fire? Stress-induced changes in facial bloodflow might betray the guilty.



Mind reading: functional MRI scans reveal brain regions that light up only during a lie.

will be ready for field testing in the next few months, Ryan says.

Another favoured technique is an infrared scanner that looks for subtle changes in capillary blood flow to the face, a sign of stress. In preliminary work published in *Nature* two years ago⁶, researchers from the DoDPI and Honeywell Laboratories in Minneapolis asked eight subjects to stab a mannequin, rob it of \$20, and then deny it. Twelve controls had no knowledge of the 'crime'.

An infrared-wielding interrogator spotted 75% of the lying muggers, and only pegged 10% of the honest controls as liars. Similar results were obtained with a traditional polygraph test. Development has continued apace, and a prototype scanner resembling a photo booth will enter its first field test this summer, Ryan says, possibly at an embassy.

The infrared scanner is easier to use than the polygraph and so the potential for its widespread use — and abuse — is higher. Without the air-conditioned booth, which maintains a constant temperature but is not essential, the scanner could be adapted for mass screening. Ryan emphasizes that there are no plans to test the scanners in airports or other US entry points. But the potential to monitor people's faces as they answer questions — "Did you pack that bag yourself?" — makes such a tool very attractive.

High stakes

Moving this technology too quickly from the lab to the field alarms experts such as Stephen Fienberg, a statistician at Carnegie Mellon University in Pittsburgh and head of the academy polygraph panel. In his experience, far too little testing is done on lie detectors. Even the *Nature* paper overstates the evidence, Fienberg argues, because it omitted ten of the original research subjects without explanation.

Lead author Ioannis Pavlidis, now a computer scientist at the University of Houston, confirms the subjects were excluded, but claims there are technical explanations. For instance, eating lunch just before the test made the recordings hard to interpret. They wrote a paper explaining this, he says, but the *Nature* format did not have enough space.

Even so, Fienberg and other experts believe researchers should emphasize new technologies' limitations, because people tend to trust machines, sometimes more than their own judgement. "If people believe these devices work, then spies or terrorists who pass through them are deemed to be OK," Fienberg says. The risk of false incrimination or false security must be taken seriously.

Another key problem with most lie-detection research is a failure to acknowledge the difference between taking part in a laboratory experiment and trying to get away with murder. Paul Ekman, a behavioural psychologist at the University of California, San Francisco, says the stakes involved make an enormous difference to a person's behaviour when lying. But the stakes are low in most experiments.

Nearly everyone agrees that more work is needed: the question is by whom. As the National Academy of Sciences report concluded, government agencies seeking to deploy lie detectors are not likely to be thorough or unbiased, nor are companies out to make a profit. "There needs to be a separation between the mission agencies who use the tools, and the people who do the core research," Fienberg says.

Ekman is one of the few doing basic research into deception. Supported by the National Institutes of Health, he has spent 30 years looking for behavioural cues that give away deception. He has found that a subtle pause in speaking or a shift in posture may indicate lying. It's not a Pinocchio's nose, he says — "but in demeanour there are signs that you are not getting a full account."

Ekman simulates high stakes for his subjects by promising \$100 or more to those who successfully deceive an interviewer and a punishment of a few hours in a dark room with sudden loud sounds if they are caught lying. Under such conditions, he has reported that trained observers can spot lies about 80% of the time⁷.

Ekman's latest project asks whether even stronger motivation can mask deception. With funding from the Defense Advanced Research Projects Agency, he is testing his approach on extremists involved in animal rights or stopping abortion. "We are not

interested in the moderate ones," he says. Like terrorists, such people may feel justified in breaking the law and lying about their activities, thus making their lies harder to detect.

While most studies measure deception indirectly, some scientists are starting to look for signs of deceit in the brain. One group, led by Daniel Langleben, a psychiatrist at the University of Pennsylvania, is using functional magnetic resonance imaging (fMRI) to look for brain regions activated by lying. Subjects received envelopes containing a playing card — the five of clubs — and \$20. With their heads inside an fMRI scanner, they were told they could keep the money if they managed to deceive a computer about what card they had. Each time they lied about having the five of clubs, two brain regions lit up, but not when they said truthfully that another card was not theirs⁸.

Human intuition

The work is extremely preliminary, Langleben cautions. Further experiments are under way to rule out alternative explanations. At this stage, fMRI is a research tool, not a way to spot liars.

Meanwhile, the Department of Homeland Security is consulting researchers on several approaches. Among them are ERPs, fMRI and voice-stress analysis, according to Gary Strong, who is in charge of behavioural studies at the department's Science and Technology division. Such tools may ultimately aid airport security and border control agents, he says. "The urgency is certainly there," Ryan agrees. "All of these technologies were in development before 9/11, but now we have an added mission."

Technology may be at the forefront of the push for new ways to detect deception, but a machine has yet to be developed that makes the human interviewer redundant. Ekman and his colleagues have developed a week-long training course in analytical interviewing which has seen a surge in interest among governmental counterintelligence groups. He has also begun to train embassy officials in the Foreign Service.

New methods will certainly come into use, both in the courts and in the ports. Whether justice and security will gain depends on whether there is sufficient evidence for their reliability. Without proper research, the question remains: who's fooling whom? ■

Jonathan Knight writes for *Nature* from San Francisco.

1. National Research Council *The Polygraph and Lie Detection* (National Academies Press, Washington DC, 2003).
2. Farwell, L. A. & Donchin, E. *Psychophysiology* **28**, 531–547 (1991).
3. Farwell, L. A. & Smith, S. S. *J. Forensic Sci.* **46**, 135–143 (2001).
4. Miyake, Y., Mizutani, M. & Yamahura, T. *Polygraph* **22**, 131–149 (1993).
5. Rosenfeld, J. P., Soskins, M., Bosh, G. & Ryan, A. *Psychophysiology* **41**, 205–219 (2004).
6. Pavlidis, I., Eberhardt, N. L. & Levine, J. A. *Nature* **415**, 35 (2002).
7. Ekman, P. *Social Res.* **63**, 801–817 (1996).
8. Langleben, D. D. et al. *NeuroImage* **15**, 727–732 (2002).

Intersex surgery disregards children's human rights

Infancy is too early to take an irreversible step that may assign a child to the wrong sex.

Sir— The Androgen Insensitivity Syndrome Support Group of Australia congratulates *Nature* on its detailed and thought-provoking article “The most important sexual organ” (*Nature* 427, 390–392; 2004). Clinical management of children affected by intersex conditions is very difficult and controversial, largely because of the many ethical and legal considerations. (See www.vicnet.net.au/~aissg for more information.)

Countries across the world, along with the United Nations, have long recognized

the rights of children to physical integrity and have banned the practice of female genital mutilation. True, the operations performed on children with intersex conditions are set in a proper clinical environment and are intended to help them develop a gender identity. But the lasting effects of reducing potential for full enjoyment of sexual experiences are often ignored — along with a person's right to make informed decisions.

What about the 8% of children with intersex conditions who are raised in the

wrong sex? Are these children's lives not worth the price of waiting to perform irreversible surgery?

The day will soon come when doctors will be sued for performing these non-therapeutic operations, and I will welcome that day. If common sense is not enough, it will take the risk of litigation to make some doctors rethink the treatment of children with intersex conditions.

Tony Briffa

AIS Support Group Australia, PO Box 1089, Altona Meadows, Victoria 3028, Australia

Dedication put Møller ahead, not fabrication

Sir— As collaborators of Anders Pape Møller we were shocked and surprised to read that he could have been accused of data fabrication (“Prolific ecologist vows to fight Danish misconduct verdict” *Nature* 427, 381; 2004).

We have never had cause to be concerned about any aspect of our collaborations. Møller is an admirable scientist, and his great organizational skills are a model of how to be productive in the face of competing time demands.

Most people are capable of much more than they actually accomplish, but they lack Møller's dedication and self-discipline. This is the secret of his phenomenal success, which has been so puzzling to the rest of the community. Clearly, these achievements have caused negative responses from some of his competitors. We would like to see a full, objective and independent inquiry into the allegations made against Møller.

Our experience tells us that Møller has the ability to completely focus on any task at hand, be it fieldwork, data analysis or writing papers. This, combined with more than a little natural talent, is sufficient to explain his exceptional productivity. We have worked with Møller on several projects, including collecting data under sometimes arduous conditions, and in all our dealings with him his behaviour has been beyond reproach.

We would ask colleagues to refrain from further public condemnation until any allegations have been proven beyond doubt.

Juan Moreno*, Tim Mousseau†

**Departamento de Ecología Evolutiva, Museo Nacional de Ciencias Naturales-CSIC, J. Gutiérrez Abascal 2, E-28006 Madrid, Spain*

†Department of Biological Sciences,

University of South Carolina, Columbia, South Carolina 29208, USA

Signed on behalf of 31 other international co-authors, whose names and contact details are available at <http://cricket.biol.sc.edu/dedication/authors.html>.

Getting in a twist again

Sir— I must start by saying that I am somewhat anal retentive (in other words I am a scientist), although I prefer the personality descriptor ‘analytical’. That said, I could not help but note that a chromosomal DNA duplex used to illustrate a recent News and Views article (*Nature* 427, 593; 2004) has a left-handed screw axis. I note the occurrence of these not-infrequent aberrations as a form of personal entertainment with perhaps a touch of masochism, as it annoys the hell out of me for some reason.

Anyway, this example was particularly noteworthy because of the Correspondence by Stanley Scher in the same issue (*Nature* 427, 584; 2004) in which he rebuts an earlier comment by John Maddox about the Watson and Crick duplex structure being self-evident. Apparently it is not. It seems that even after 50 years and countless renderings of the famous B-form DNA duplex, people still get it wrong about 15% of the time (my non-peer-reviewed, unpublished and incomplete survey). I can only conclude that the structure and its correctness are anything but self-evident, at least as far as advertisers and scientific illustrators are concerned.

In fairness I must note that there is the outside chance that Figure 1 was depicting a rare and controversial chromosomal Z-DNA domain, but even so, the helical parameters are way off.

Eric “my DNA is (mostly) right-handed and double-stranded” Henderson

Genetics, Development and Cell Biology, Iowa State University, Ames, Iowa 50011, USA

US visa restrictions harm job prospects abroad too

Sir— It was with great interest that I read your News Feature “As one door closes...” (*Nature* 427, 190–195; 2004), on restrictive visa requirements in the United States and the impact they are likely to have on science. As a postdoctoral researcher from Australia, working in Alabama, I recently discovered how US visa restrictions can affect career choices for people already working in the United States.

Last June, after working in the United States for about five years, I submitted an application to renew my H1B visa. In February this year, the Immigration and Naturalization Service (INS) finally began to look at my application. In the meantime I had applied for an academic position in my native Australia. To my delight I was granted an interview; this was a dream job for me. But after applying for a US visa, you cannot leave the United States without jeopardizing your application. So I was faced with a tough choice: to go to the interview and risk losing my current job, or to stay in the United States and miss the job opportunity. With employment prospects as limited as they are in the natural sciences, I decided to travel to the interview.

A few weeks ago I was told that the INS would not be considering my visa renewal, and I would have to leave the country. Fortunately for me the gamble paid off, as I got the job in Australia. Nonetheless, I will have to cut my research short in the United States and, had I not got the job, I would now be unemployed. In my view, the way that the INS can effectively stop you pursuing jobs for protracted periods will make the United States even less appealing to foreign postdocs and academics.

Toby F. Bolton

Dauphin Island Sea Lab, 101 Bienville Boulevard, Dauphin Island, Alabama 36528, USA

Earth's crude mosaic

How plate tectonics triggered a seismic upheaval in geology.

The Earth: An Intimate History

by Richard Fortey

HarperCollins: 2004. 502 pp. £25. To be published in the US by Knopf in November.

Gordon L. Herries Davies

Neither names nor place will I reveal, but the year was 1968. We had just heard a lecture delivered by a geological bigwig of the day. His subject had been the new geological synthesis, and he had preached the gospel of plate tectonics. As we left the auditorium, one of my colleagues — a noted palaeontologist — commented: “He really seems to take all this quite seriously!”

My colleague was not unique. We were living at the time amidst a revolution in geological thought. Our intellectual world was experiencing massive seismic upheaval. Those of us raised on the first edition of Arthur Holmes' *Principles of Physical Geology* had once marvelled at the daring that had allowed him to devote a whole final chapter to continental drift. Yet by the 1960s we were being told that our continents ride mobile plates as performers ride their bucking mounts in some Western rodeo. Exciting stuff. I feel it a great privilege to have been an Earth scientist during a period when geology sprang back into vigorous life after well-nigh a hundred years of semiparalytic lethargy.

An eminent palaeontologist, Richard Fortey has achieved an international reputation as an expositor of the Earth sciences, as the cover of his latest book emphasizes: his name appears in the same typeface as the title. In *Life: An Unauthorised Biography* (HarperCollins, 1997), Fortey traced the natural history of the Earth over four billion years. Now he has turned his attention to the annals of the very Earth itself. The tale has been largely rewritten since 1963, when Fred Vine and Drummond Matthews published in *Nature* their famed essay “Magnetic anomalies over ocean ridges”.

Fortey takes my mind still further back, to the swinging big bands of my youth. In the opening number I applaud when, under Fortey's direction, the ‘bones resume their seats and the golden sax section takes over. The star trumpet inspires foot-tapping ecstasy with an up-tempo eruption. The male vocalist takes over the mike to add his gravelly contribution. And all the while the entire outfit swings forward to the exciting, thrusting beat of the rhythm section. Count Basie and Benny Goodman may have gone, but the sheer energy of Fortey's plate-tectonic rhythm section ensures that the geology big band will continue to swing before large audiences for decades to come.



On the trail: a trip down the Grand Canyon's stratigraphy is like reading a 'diary of the Earth'.

I wholeheartedly approve of Fortey's literary strategy. He wishes us to remember that his is a field-based science. Through these pages we are taken to examine actual rocks. One illustration facing page 214 says it all — five geologists confront a classic section. In days of yore they would all have sported hammers, but today we are in conservation mode and not a single hammer is to be seen.

The book is constructed around a framework of visits to specific geological locations scattered around the world. For each location, Fortey offers a clear, graphic and entertaining exposition of the manner in which, over an eon, the observed geological phenomena have achieved their present state. And he forcefully reminds us that events remotely embedded in deep time may yet be highly relevant as determinants for the lifestyles of modern human communities. And all the while, on page after page, behind Fortey's scintillating solo breaks, we hear the steady, driving beat of that plate-tectonic rhythm section.

The first location to which Fortey takes us is the Bay of Naples, to imbibe the lessons of Vesuvius, Pompeii and the columns of the Temple of Serapis. He finds other didactic locations in Dartmoor and Scotland's

northwest Highlands, in the Czech Republic and the Alps, in the Indian Deccan traps and Hawaii, in California and Newfoundland. I have had the good fortune to visit many of Fortey's locations, and his handsomely crafted chapters evoke abundant memories. A traverse of the exposed plane of Scotland's Moine Thrust was for Ben Peach so emotional an experience that it commonly brought tears to his eyes. I feel much the same about the Grand Canyon, which also features here. Fortey terms the sequence exposed there “the diary of the earth”. Never again will I hear the third movement (‘On the Trail’) of Ferde Grofé's *Grand Canyon Suite* without thinking of Buttermilk, the mule that carried Fortey down that magnificent stratigraphic ladder to the great river itself.

Everybody who has ever written a book knows that when the advance copy arrives it instantly falls open to reveal something that one wishes had been otherwise. For Fortey this must have been the NASA photograph on page 462. This is not the Gulf of Aden, as the caption claims, but the Gulf of Aqaba. I offer just one other trivial correction. Throughout the book, Fortey pays repeated and well-merited tribute to *Das Antlitz der Erde* by Eduard Suess. On page 26 he ascribes

BUDDY MAYS/CORBIS

the English translation of this classic to Hertha Sollas, 'wife' of William Johnson Sollas — William did have two wives (not simultaneously), but Hertha was one of his two daughters. She was one of a talented group of ladies who studied science in Cambridge around 1900 but whose gender denied them any of the university's degrees.

The vast majority of publishers' blurbs are not to be taken seriously, but I think there may be some truth in the claim that this book "will change the way you view the world — permanently". This is a work that I admire and recommend. My palaeontological colleague of 1968, although long in retirement, remains scientifically active. I will certainly be telling him how much I enjoyed my tour of Earth's crude and stitched mosaic in Fortey's excellent company. ■

Gordon L. Herries Davies is at Trinity College Dublin and Ballinacough House, Nenagh, County Tipperary, Ireland.



FLIP NICKLIN/MINDEN

On the scent of a mystery: could the sperm whale's giant nose be used for sound production?

Whales with a nose for culture

Sperm Whales: Social Evolution in the Ocean

by Hal Whitehead

University of Chicago Press: 2003. 456 pp. \$80, £56 (hbk); \$30, £21 (pbk)

M. Wahlberg

More than thirty years have passed since Alfred Berzin wrote his seminal monograph *The Sperm Whale* (Israel Program for Scientific Translations, 1972). This book was based mainly on anatomical data obtained from whaling. Now, in *Sperm Whales*, Hal Whitehead enthusiastically reviews what we have learned since then about the social behaviour of these giants, owners of the planet's largest nose and brain.

Early accounts of sperm-whale behaviour came from whalers in the nineteenth century. During the revival of sperm whaling in the 1950s, several biologists, including Berzin, made a range of anatomical studies. The whales' diet was studied by analysing stomach contents, and their breeding behaviour was inferred from sexual dimorphism and the anatomy of the reproductive organs. For practical reasons, there seemed to be no other way of studying this large, deep-diving, offshore whale.

In the early 1980s, this view changed dramatically when Whitehead and Jonathan Gordon, both then at the University of Cambridge, UK, led the *Tulip* expedition to Sri Lanka and showed that whaling was not the only way to study this species. Population data could be gathered using photo-identification and acoustic techniques, and long-term behavioural studies gave intriguing

insights into the social lives of sperm whales.

Whitehead's book is a fascinating account of what has been learned during two decades of non-lethal research inspired by the *Tulip* expedition. The core of sperm-whale society is now known to consist of female groups living in warm waters. When males grow up, they split from these groups and migrate towards polar waters to feed. As adults, they return to the tropics to rove among the female groups. By contrast, the females form decade-long bonds with other whales within their group. These bonds are not immediately obvious to a casual observer but have been revealed through long-term field studies.

An important part of the book deals with the animal's nose. This huge fat-container, plumbed with fire-hose-sized nasal passages and air sacs, is surrounded by a muscle layer so thick that it is possible for the whale to lift a truck with a frown, and it is innervated with nature's thickest nerve. Whitehead dismisses several of the existing theories as to its function, such as a buoyancy control mechanism or a battering ram. Instead, he favours recent acoustic work showing that the whale uses its nose for sound production, thereby associating the largest nose on the planet with the loudest sounds produced in the animal kingdom.

Sound is fundamental to the lives of sperm whales. The book summarizes the results from several studies of their acoustic repertoire. The whales use knocking sounds both for communication and to find food in the abyssal darkness. The animals use their communication sounds, called codas, to maintain contact with one another. Sympatric groups of sperm whales have different coda repertoires, and these, along with other behavioural differences, seem to be

culturally transmitted between generations.

Whitehead's research group at Dalhousie University in Nova Scotia is spearheading work on cultural transmission and the social evolution of this species. However, some of the conclusions drawn from their studies seem premature, considering the limited amount of data so far available. The book does not discuss the serious scepticism expressed by some scientists about these conclusions (see *Behav. Brain Sci.* **24**, 309–382; 2001).

These few, small criticisms aside, the book is a well-written, comprehensive and much-needed sperm-whale monograph, and includes a detailed bibliography covering all aspects of current research. Whitehead's survey also reveals that many aspects of this species' behaviour remain unknown. Unlike the whalers of the 1800s, today's scientists approach the animals in a peaceful manner to find out how they are related and where and how they catch their prey. The rapid developments in genetic sampling methods and data logging mean that these techniques also can be applied to the whales. Instead of using harpoons, today's scientists use darts to obtain skin samples, and attach suction-cups linked to instruments for sampling the animals' acoustic and diving behaviour.

Whitehead's book will serve as an important source of inspiration and information for whale scientists. The book is also entertaining for the general reader, and is suitable educational material for students of marine sciences. ■

M. Wahlberg is in the Department of Zoophysiology, Aarhus University, DK-8000 Aarhus C, Denmark, and Tjärnö Marine Biological Laboratory, Department of Marine Ecology, Göteborg University, Sweden.

Statistically unlikely

Francis Galton: Pioneer of Heredity and Biometry

by Michael Bulmer
Johns Hopkins University Press: 2003. 376 pp.
 \$45, £33.50

Eileen Magnello

The life of the Victorian polymath Francis Galton — who explored South Africa, created weather charts, introduced finger-printing, studied anthropology, devised statistical methods and is perhaps best known for fostering ideas of hereditary improvement of humans through eugenics — continues to exert fascination for historians and scientists. A recent biography by the geneticist Nicholas Wright Gillham (see *Nature* **415**, 19–20; 2002) is now joined by statistician Michael Bulmer's book, which provides a distillation of some of Galton's ideas of heredity and statistics, reinterpreted from current thinking.

Bulmer's writing is very present-minded — many of Galton's ideas are taken out of the Victorian context in which he lived and worked. Instead, Bulmer imagines how Galton would have dealt with such problems if he were alive today, even though Galton could not have outlined a theory of quantitative genetics, as the discipline did not exist in his lifetime; nor could he have had any discussions on 'phenotypic variance', as this is an expression introduced by R. A. Fisher in the 1920s.

Bulmer thus provides little understanding of how Galton and his contemporaries would have tackled problems of heredity or statistics. When he does consider some of Galton's Victorian ideas, he simply construes that they reveal more about the opinions of Galton's contemporaries "than about the real world" — as if to say that Galton's contemporaries were not living in the real world. The reader thus never learns how the philosophical ideologies underpinning Galton's thinking, or that of the Belgian statistician Adolphe Quetelet, influenced their ideas about the normal distribution. Bulmer can only conclude that "Galton was mistaken to adhere to Quetelet". Yet Quetelet attached so much significance to the normal curve because of his belief in determinism, whereas Galton's belief in essentialism — which was the dominant thinking of the taxonomists, typologists and morphologists until the end of the nineteenth century, and gave rise to the morphological concept of species — implied that species regressed to the mean value. Galton was, therefore, convinced that all biological data could only be normally distributed.

The problem with Bulmer's 'internalist' approach is that the reader never really knows what Galton tried to do, as Bulmer

Website

Bringing Galileo to life

The revamped website of the Institute and Museum of the History of Science (IMSS), housed in one of the oldest palaces in Florence, the Palazzo Castellani, will be as valuable in its own way as the museum's collection.

The physical collections are heir to five centuries of acquisitions by the powerful florentine Medici and Lorraine families. They include innumerable scientific instruments and anatomical models — as well as Galileo's finger, which is famously displayed alongside many of the instruments he used (for a fuller description see *Nature* **425**, 128; 2003). They are among the most important collections in the world.

The new website animates the individual instruments, as well as describing and cataloguing them. For example, it describes how



Galileo's compass came to be developed (see picture), and shows, through clever simulations, how it actually works.

The website, launched on 24 March, is a work in progress, which will bring more life to more instruments — and to the museum itself — in the coming years.

Alison Abbott

www.imss.fi.it

wants to tell us instead what Galton should have done. Given Galton's own limitations with mathematics — he only managed to get a third class in the mathematics tripos at Cambridge University and had a breakdown during this time — he would have found Bulmer's mathematics impossible to understand, as, indeed, will anyone who is not a professional statistician.

Like his cousin Charles Darwin, Galton worked as an independent scientist and never held a university post. Inspired by Darwin's ideas on biological variation, Galton began to devise a statistical method to measure it. His statistical work caught the attention of the zoologist W. F. R. Weldon, who was looking for a way to find a working hypothesis for Darwin's variation. In 1891, Weldon took up the Jodrell chair of zoology at University College London, where Karl Pearson was professor of applied mathematics. In 1892, Pearson, who had just devised the standard deviation, struck up an intellectual partnership with Weldon, who needed help interpreting a bimodal distribution (at a time when it was thought that all data had to conform to the normal distribution). Weldon introduced Pearson to Galton in 1894, but it was Weldon who gave Pearson the impetus he needed to develop a new statistical methodology. Bulmer's discussions on Weldon are entirely devoid of that vital interaction he had with Pearson; Bulmer thus fails to show that the reason Weldon was the first scientist to provide empirical evidence of natural selection was precisely because Pearson was devising new statistical tools for him. Although Pearson grew fond of Galton, this relationship did not really develop until after Weldon's premature death in 1906.

Pearson has long been erroneously viewed as a disciple of Galton who followed

in his footsteps and merely expanded what Galton started — a view that Bulmer endorses. Thus, Bulmer's interpretation of galtonian and especially pearsonian statistics is deeply problematic. It is wrong to assume that the impetus to Pearson's statistics came from reading Galton's *Natural Inheritance* (Macmillan, 1889), especially as Pearson's initial reaction to this book was actually quite cautious. Pearson did not incorporate Galton's statistical ideas on correlation and regression for another six years, when Galton wrote to him asking for his assistance in 1895. By then, Pearson had already created the infrastructure of his statistical methodology. It was not until 1934, when Pearson was 78 years old, that he reinterpreted in a more favourable light the impact that Galton's book had on his own statistical work.

There is a long discussion of Galton's idea of correlation and statistical regression, but Bulmer never explains that Galton actually confused correlation with statistical regression. The reason the correlation coefficient is referred to by the letter *r* and not, say, *c* is that Pearson showed that Galton's correlation formula was a measure of regression instead, as it measured the slope of the regression line, and he retained Galton's *r* to symbolize the correlation coefficient.

This book is not so much a story about Galton as it is about Bulmer recalculating Galton's data using twentieth-century statistical methods. Perhaps Bulmer should have simply taken all of Galton's data and reanalysed it using contemporary statistical methods to gain a different understanding of Galton's data.

Eileen Magnello is at the Wellcome Trust Centre for the History of Medicine at University College London, 24 Eversholt Street, London NW1 1AD, UK.

The productivity of failures

How a rejected paper generated a flourishing research programme.

Ernst Fehr

In the late 1980s, I wrote a paper about involuntary unemployment, which I hoped would answer questions economists had been debating for decades. The paper was rejected, but developed into basic research that encompassed several disciplines — biology, psychology and economics.

Involuntary unemployment is a state in which unemployed workers are willing to work for less than the going wage, but nevertheless cannot get a job. Common sense dictates that in periods of economic downturn, some workers without a job are involuntarily unemployed. In economics, there is a long-standing question as to whether involuntary unemployment can also occur in a state of equilibrium, that is, when no endogenous economic forces push for a change in the situation. A more complex and politically contested issue is whether an equilibrium with involuntary unemployment can occur in a free-market economy that is unconstrained by union power and minimum-wage legislation. The great British economist John Maynard Keynes argued that this is possible, but economists have grappled with this question for decades — and still remain divided.

My paper in the 1980s was entitled 'Fair wages and unemployment', and I thought it provided a convincing answer to these problems. Its central theme was that employees do not work hard if they are paid an unfair wage. Therefore, if the going wage represents a fair standard for reference, workers hired below this wage will be willing to accept an initial low wage offer, but are later — when employed — less likely to work hard. Hence, employers shy away from cutting employees' wages or from employing workers below the going wage.

This hypothesis, that fairness concerns affect the functioning of the labour market, is an old one, but mainstream economics has largely disregarded this idea. Just like other people, most economists appreciate being treated fairly in their personal interactions — and would be deeply unhappy if paid less than what they consider to be fair. But when it comes to the modelling of economic affairs, fairness is not a currency that counts. In addition (like any other science), economics is characterized by conventions, one of the strongest being that a good model should generate predictions that are based on the



assumption that all agents maximize well defined objective functions. In principle, it is not that important what economic agents maximize, as long as they maximize something. But in practice, economists still have strong conventions regarding what constitutes a proper objective function — fairness has certainly not been among the objectives the convention legitimizes. Thus, I was not too surprised when several leading economics journals rejected my paper, although it was built on a framework where all agents maximize 'something'. Unfortunately, the 'something' also included fairness goals.

I remember that one of the referees argued that the distance between the assumptions and the conclusions was "too small" to make the paper interesting. This argument reveals a lot about the psychology behind judging a paper's merits. A paper has value when it creates a surprise — either empirically or in terms of theoretical insights. Science is often the prisoner of the psychology of surprise, and for theorists, elegance or mathematical beauty is also important. Despite all the rejections, I remained convinced that fairness concerns are important. In the end, it is not important whether a theory is simple or complicated, elegant or clumsy, beautiful or ugly — such considerations are of a secondary nature, alien to true science. Usually they are just a poor substitute for a simple fact — the absence of convincing empirical evidence. Therefore, I searched for ways to capture the issue of fairness empirically. In retrospect, this search was the major turning point of my scientific career and completely reshaped my research programme. I metamorphosed from a labour-market theorist into an experimental economist.

In the late 1980s there was already a large amount of experimental literature on fairness in bilateral bargaining, which started

with the economist Werner Güth's path-breaking ultimatum game experiments. However, I was interested in whether fairness motives affect competitive markets. Experimental economists had already conducted thousands of competitive market experiments and the conventional wisdom that fairness is irrelevant in markets prevailed. But typically these experiments differed from the environment in my rejected theory paper because effort was assumed to be fully determined in the contract. Fortunately, I received a grant

enabling me to employ two excellent research assistants — Georg Kirchsteiger and Arno Riedl. Together, we designed an experiment which exactly implemented the economic environment assumed in my theory paper — in particular, that of non-contractible effort. We almost could not believe the data from the pilot experiment — we saw what we had only hoped for in our wildest dreams: fairness concerns strongly inhibited competitive forces and kept wages substantially above the competitive level. Moreover, as our theory predicted, wage reductions induced the employees to reduce their effort, rendering wage cuts unprofitable for the employers. Now we had the surprise we needed for publication.

My later work, and that of many others, indicated that fairness concerns permeate many other aspects of economic and social life — they affect competition and cooperation between and within firms, they influence international negotiations, the provision of public goods, exploitation of common property resources and they are the basis of many political conflicts. In addition, fairness norms probably played a decisive role in the evolution of human sociality. What began as research on involuntary unemployment turned into basic research about the nature of human altruism (see *Nature* 415, 137–140; 2002, and *Nature* 425, 785–791; 2003) and led to the concept of strong reciprocity — a term invented by Samuel Bowles and Herbert Gintis to describe the widespread human propensity to reward helpers and to punish cheaters. Today, this concept inspires anthropologists, zoologists, evolutionary biologists and evolutionary psychologists alike. I wish them success, and a few productive failures. ■

Ernst Fehr is at the University of Zürich, Institute for Empirical Research in Economics, Blümlisalpstrasse 10, CH-8006 Zürich, Switzerland.

CORBIS

Lost and found

Neil H. Shubin and Randall D. Dahn

Can we ever hope to pin down the genetic changes that underlie the big steps in evolution? Possibly so, if a study of the variation in the pelvic fins of sticklebacks is anything to go by.

Darwin's lament¹ that "Our ignorance of the laws of variation is profound" has described one of the persistent problems in evolutionary biology for the past 145 years. How does genetic variation — the raw material of evolution — arise within populations, and how does it evolve to make species anatomically and behaviourally distinct? Ultimately, answers to these questions will be derived from understanding the actual genetic changes that control the evolution of anatomy or behaviour. These insights could explain a host of puzzles, from how rapid evolutionary change can occur in populations, to why morphological evolution often follows certain well-defined trends. On page 717 of this issue, Shapiro *et al.*² describe the genetic basis for one such trend: limb reduction in threespine stickleback fish.

Surprisingly, some of the most significant novelties in the history of life are associated not with the evolution of new structures but with the loss or reduction of primitive ones. In vertebrates, for example, the invasion of new ecological niches and the origin of new locomotor adaptations involve either the complete loss or partial reduction of appendages^{3,4}. The complete loss of appendages has been involved in the evolution of new aquatic lifestyles in whales and burrowing niches in snakes, amphisbaenians (worm lizards) and caecilians (rubber eels). The reduction or loss of fingers and toes is associated with the evolution of jumping, flying and running in some amphibians, reptiles and mammals. In sticklebacks (Fig. 1), a large spine in the pelvic fin serves to limit predation in marine populations. But in some freshwater stickleback populations this pelvic spine is reduced or lost entirely — a trend that might be associated with lower concentrations of available calcium ions in freshwater lakes, or with the presence of large invertebrates that feed on sticklebacks by grasping the pelvic spines.

By comparing the genomes of sticklebacks with and without pelvic structures, Shapiro *et al.*² have identified the major gene responsible for the variation in pelvic size in these populations of threespine sticklebacks. By crossing a female from a population that has a complete pelvic skeleton with a male from a population that lacks pelvic spines, more than 375 progeny were produced after two generations. The offspring showed a range of conditions, from complete pelvic fins to reduced or wholly absent pelvic skeletons.



Figure 1 Sticklebacks under scrutiny. The threespine stickleback, *Gasterosteus aculeatus*, is shown in ventral view, with bony structures stained red. The right-most group of bones makes up the pelvic-fin structure under discussion here². Inset, sticklebacks *au naturel*.

By correlating how fin morphology segregates with each of 221 genetic markers in the offspring, Shapiro *et al.* showed that five regions of the genome are associated with pelvic reduction. One region accounted for a large amount of the pelvic reduction in freshwater populations; the other four explained smaller amounts of the reduction. Importantly, the *Pitx1* gene maps precisely to the site that is associated with most of the variation. This is strikingly consistent with data from mice: mutants lacking *Pitx1* function also have reduced hindlimbs, and do so asymmetrically, with the right limb more affected than the left^{5–7}. Indeed, this same asymmetry is seen in the natural variation in pelvic fins within freshwater stickleback populations.

Pitx1 mutations in mice are often lethal, because they cause developmental abnormalities of the head, face and some glands. How, then, could alterations in this gene be involved in limb reduction in living populations of stickleback fish? The answer is that the regulation of *Pitx1* — not the protein encoded by the gene — has changed. Thus, Shapiro *et al.* found that the sequence of the protein-coding region of the *Pitx1* gene is identical between the different populations of sticklebacks. But the gene's expression pattern is altered markedly: the population with complete pelvic loss shows no *Pitx1* expression in

appendages but retains patterns of gene activity in other areas, such as the thymus, olfactory pits and caudal fins (Fig. 2, overleaf). This type of localized decrease in the activity of *Pitx1* can result in pelvic-fin reduction without affecting other parts of the body.

Regulatory changes affect when and where a gene is active, not the actual product of the gene. So these types of changes are often involved in non-lethal and rapid morphological change, and are likely to be extraordinarily important components of evolutionary history^{8,9}. Indeed, stratigraphic and geographic analyses suggest that limb loss in sticklebacks has evolved in fewer than 10,000 generations². Extrapolating these results to other taxonomic groups leads to the conclusion that major morphological change can evolve rapidly through regulatory changes in a small number of genes. In addition, these changes can happen independently in different populations and species. Systematists have long known that some evolutionary changes occur more readily than others — different species often evolve similar traits independently. Shapiro *et al.* have shown that such patterns of parallel evolution might have the same genetic basis: similar genetic shifts seem to underlie pelvic reduction in stickleback populations in Canada and Iceland, more than 5,700 kilometres apart. Indeed, these results might speak to general principles

M. D. SHAPIRO *ET AL.*

M. PETERSON/ARDEA

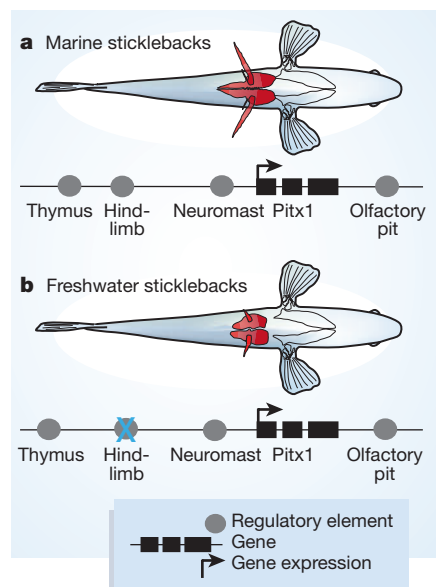


Figure 2 Evolution in action: a genetic basis. Shapiro *et al.*² have found that variation in expression of the *Pitx1* gene underlies variation in reduction of a large spine in the pelvic fin in threespine sticklebacks. The regulation of *Pitx1* activity has been partitioned into discrete regulatory elements, each of which controls this gene in a particular tissue; changes in *Pitx1* expression that result in pelvic-fin reduction can thus be uncoupled from the requirements of other parts of the body for *Pitx1* activity. a, Shapiro *et al.* show that marine sticklebacks with intact *Pitx1* regulatory elements develop robust pelvic spines. b, Mutations (X) that specifically affect hindlimb *Pitx1* expression alter the pelvic-fin structures of freshwater species, whereas other *Pitx1*-dependent structures are unaffected. The neuromast is a mechanosensory organ. Figure adapted from ref. 2.

of parallel evolution, because examples are now cropping up from diverse places. In bird plumage¹⁰ and stickleback armour¹¹, the independent evolution of similar features in different populations appears to involve similar genetic modifications.

One of the central mysteries of evolutionary biology has been the relationship between microevolution and macroevolution¹². How can an understanding of the evolutionary mechanisms that act in populations today explain the types of variation that distinguish higher taxonomic groups, such as genera, families or even phyla? Can an understanding of population-level processes explain major evolutionary events such as the Cambrian explosion — the period around 550 million years ago when complex animal life took off?

Perhaps so. Shapiro *et al.*² might have discovered a smoking gun — a real example of a type of macroevolutionary change that is produced by genetic differences between populations. ■
Neil H. Shubin and Randall D. Dahn are in the Department of Organismal Biology and Anatomy, University of Chicago, Chicago, Illinois 60637, USA. e-mail: nshubin@uchicago.edu

1. Darwin, C. *On The Origin of Species* (Murray, London, 1859).
2. Shapiro, M. D. *et al.* *Nature* **428**, 717–723 (2004).
3. Morse, E. S. *Ann. Lyceum Nat. Hist. NY* **10**, 141–158 (1872).
4. Lande, R. *Evolution* **32**, 73–92 (1978).
5. Lanctôt, C. *et al.* *Development* **126**, 1805–1810 (1999).
6. Szeto, D. P. *et al.* *Genes Dev.* **13**, 484–494 (1999).
7. Marcil, A. *et al.* *Development* **130**, 45–55 (2003).
8. Carroll, S. B. *Cell* **101**, 570–580 (2000).
9. Levine, M. & Tjian, R. *Nature* **424**, 147–151 (2003).
10. Mundy, N. I. *et al.* *Science* **303**, 1870–1873 (2004).
11. Colosimo, P. F. *et al.* *PLoS Biol.* doi:10.1371/journal.pbio.0020109
12. Gould, S. J. *The Structure of Evolutionary Theory* (Harvard Univ. Press, Cambridge, MA, 2002).

Astronomy

The missing black-hole link

Nate McCrady

A class of black holes of intermediate mass is expected but has never been detected. The suggestion that these beasts might lurk behind powerful X-ray sources in nearby galaxies is now strengthened.

Black holes are known to exist in two mass regimes: those between two and ten times the mass of the Sun, known as stellar-mass black holes and formed from the collapse of the most massive stars; and those between a million and a billion times the mass of the Sun, the ‘supermassive’ black holes. Such behemoths, including the one at the centre of the Milky Way¹, are the engines powering quasars and active galactic nuclei. Between the two extremes are intermediate-mass black holes, although these have never been detected unambiguously. Observations from NASA’s Chandra space observatory have stoked the debate

over their existence. The images show extremely luminous, compact X-ray sources lurking in and around star clusters in nearby galaxies — sources that might be associated with intermediate-mass black holes.

Central to the discussion is the lack of a credible mechanism by which these objects could have formed. But dynamical simulations by Portegies Zwart *et al.*, reported on page 724 of this issue², show that stellar collisions in dense star clusters could ‘run away’ and lead to the formation of an intermediate-mass black hole at the cluster core. This process is particularly intriguing because the accretion of such

black holes early in the galaxy-formation process is a crucial step in the formation of supermassive black holes at galaxy centres.

The nature of ‘ultraluminous X-ray sources’ (ULXs), identified in high-resolution Chandra images (Fig. 1), is a key question in the black-hole debate. Bright X-ray point sources in galaxies are generally either young remnants of supernovae, which fade over time, or the compact remnant of a massive star (a neutron star or a stellar-mass black hole). The latter shine as they accrete matter, typically from a companion star. The brightness of such an object generally cannot exceed the Eddington limit, which is the luminosity at which radiation pressure from escaping photons overcomes gravity and disperses the accreting material. ULXs are deemed ‘ultraluminous’ because the isotropic luminosities implied by their measured fluxes exceed the Eddington limit for the largest black holes formed from single massive stars. If ULXs are actually black holes accreting matter, then their masses must fall in the 100 to 1,000 solar-mass range — the missing intermediate-mass black holes.

Observations^{3,4} indicate that ULXs tend to be located in or near star clusters in starburst galaxies, sites of very active star formation. This connection prompted the suggestion^{5,6} that intermediate-mass black holes are formed in ‘super star clusters’ — young, dense clusters 10 to 100 times more massive than the ancient globular clusters they closely resemble. The largest stars formed in such clusters have masses 100 to 150 times that of the Sun, but they leave behind black holes of only 15 to 20 solar masses.

To form an intermediate-mass black hole, a mechanism other than star formation is required: Portegies Zwart *et al.*² propose runaway growth prompted by the collision of stars. In the field of a disk galaxy, outside clusters, collisions between stars are very unlikely, because their cross-sectional area is so small compared with the vast distances between them (galaxies are, in fact, far more common collision partners). Dense star clusters, however, are a different matter⁷. The most massive stars in a cluster tend to sink in towards the core, by transferring energy to lower-mass stars in a process known as mass segregation. Once the central density becomes high enough, collisions between high-mass stars can occur. But can these collisions create a ‘runaway star’ that builds sufficient mass to beget a 100–1,000-solar-mass black hole, before the high-mass stars explode as supernovae?

Portegies Zwart *et al.* have used many-body simulations to examine the behaviour of a pair of young super star clusters⁸ in the nearby starburst galaxy M82 (Fig. 1). One cluster, MGG 11, coincides spatially with the brightest ULX discovered so far, the luminosity of which corresponds to an intermediate-mass black hole of 300–900 solar masses⁹.

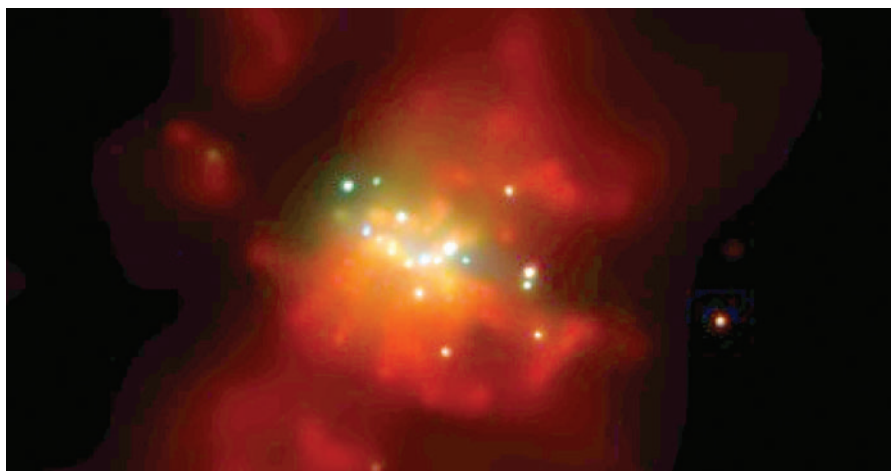


Figure 1 **Bright lights.** Captured in this false-colour image from the Chandra X-ray Observatory is a population of extremely bright, point-like sources in the galaxy M82, 11 million light years from Earth. Their 'ultraluminous' emission might result from the accretion of matter by intermediate-mass black holes. Dynamical simulations by Portegies Zwart *et al.*² suggest that in at least one star cluster in the galaxy — MGG 11, which coincides with one of the ultraluminous sources — collisions between stars could have created a 'runaway' star that ultimately became an intermediate-mass black hole.

The authors' simulations indicate that, in this cluster, a runaway star could form on the necessary timescale. The result is tantalizing — this could well be how the building blocks of supermassive black holes formed.

Deep-space images from the Hubble Space Telescope indicate that galaxy mergers were common at early times in the history of the Universe, and the conditions in the centres of those merging galaxies might have resembled those seen in nearby starburst galaxies today. The link between ULXs and intermediate-mass black holes is, however, not conclusive. There are plausible alternative explanations for the high X-ray luminosities of ULXs, such as beaming¹⁰ (the observed X-ray flux is emitted through a narrow opening angle rather than isotropically) and super-Eddington emission (attributed to 'clumpy' accretion disks¹¹). More high-resolution observations of super star clusters and ULXs, coupled with simula-

tions of star-cluster dynamics, will ensure that the debate over intermediate-mass black holes continues.

Nate McCrady is in the Department of Astronomy, University of California, Berkeley, California 94703, USA.

e-mail: nmccrady@astro.berkeley.edu

1. Ghez, A. M., Morris, M., Becklin, E. E., Tanner, A. & Kremenek, T. *Nature* **407**, 349–351 (2000).
2. Portegies Zwart, S. F., Baumgardt, H., Hut, P., Makino, J. & McMillan, S. L. W. *Nature* **428**, 724–726 (2004).
3. Kaaret, P. *et al.* *Mon. Not. R. Astron. Soc.* **348**, L28–L32 (2004).
4. Fabbiano, G., Zezas, A. & Murray, S. S. *Astrophys. J.* **554**, 1035–1043 (2001).
5. van der Marel, R. P. in *Coevolution of Black Holes and Galaxies* (ed. Ho, L. C.) (Cambridge Univ. Press, in the press).
6. Gurkan, M. A., Freitag, M. & Rasio, F. A. *Astrophys. J.* **604**, 632–652 (2004).
7. Bonnell, I. A. & Bate, M. R. *Mon. Not. R. Astron. Soc.* **336**, 659–669 (2002).
8. McCrady, N., Gilbert, A. M. & Graham, J. R. *Astrophys. J.* **597**, 240–252 (2003).
9. Kaaret, P. *et al.* *Mon. Not. R. Astron. Soc.* **321**, L29–L31 (2001).
10. King, A. R., Davies, M. B., Ward, M. J., Fabbiano, G. & Elvis, M. *Astrophys. J.* **552**, L109–L112 (2001).
11. Begelman, M. C. *Astrophys. J.* **551**, 897–906 (2001).

Cancer

Kip moving

John G. Collard

The p27^{Kip1} protein inhibits cell proliferation, helping to prevent tumours developing. We now know that it also affects cell migration, by regulating Rho proteins. Does this function influence tumour progression?

Tumours arise when cells proliferate in an uncontrolled manner and, later, take on malignant characteristics, such as the ability to invade surrounding tissues. Particularly aggressive tumours can also metastasize — they spread to distant organs. These changes in tumour-cell properties are driven by alterations that favour various steps in the process: for instance,

lower concentrations of the tumour-suppressor protein p27^{Kip1} lead to deregulated cell proliferation and so are thought to promote tumour formation. There are hints that this protein might also control other stages of cancer development; notably, it regulates cell migration — a prerequisite for metastasis. Writing in *Genes and Development*, Besson *et al.*¹ suggest how it does so.



100 YEARS AGO

The administration of chloroform is a subject that is of personal and direct interest to everyone in this present age of civilisation. Sooner or later either we ourselves or those dear to us gladly accept the relief from suffering that is offered, and that chloroform shall be given so that no unavoidable risk is run is a necessity that forces itself on our attention. That much remains to be done in the direction of safety is only too evident. We confess to perusing the diagram of the yearly increasing death-rate from chloroform... with a feeling of horror, and that is deepened when we read the instances given of such deaths, and supplemented by others which have come to our knowledge independently, where chloroform has been given for a trifling operation to an otherwise healthy patient, and where the phrase "Death from cardiac syncope" has acted as an anaesthetic to the conscience of the ignorant and careless anaesthetist.

From *Nature* 14 April 1904.

50 YEARS AGO

High Altitude Rocket Research. Men have long been attracted by the possibility of being able to carry out explorations far above the earth's surface. In 1784 Vincent Lunardi made the first balloon ascent from London, his achievement arousing great popular enthusiasm. Attention, too, was directed to the advances in knowledge which might be gained... When rather more than a century and a half later rockets began to fall on London they aroused different feelings; but men of science found some compensation in the recognition of the fact that a weapon of real use had been developed. During the early part of 1946, an important meeting was held in Washington, D.C., at which the exploitation of the German invention was discussed and a programme of fundamental research was formulated... A rapidly moving rocket does not, of course, provide an ideal platform on which to mount measuring instruments; but in spite of this many results have been derived. Perhaps the most valuable are those relating to the density and the temperature of the upper atmosphere, those relating to the electric current system causing the diurnal variation in the magnetic field, and those relating to the solar emission in the spectral range which cannot be observed from ground-level.

From *Nature* 17 April 1954.

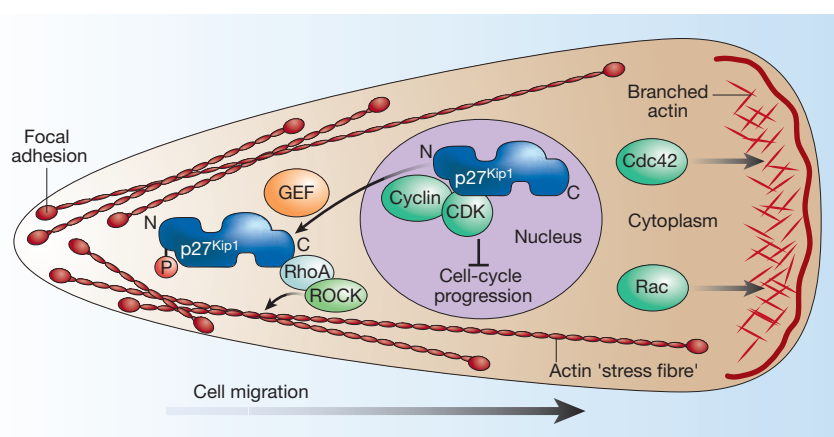


Figure 1 Dual function of a tumour suppressor. In the cell nucleus, the tumour-suppressor protein $p27^{Kip1}$ binds with its amino-terminal region (N) to complexes of cyclins and cyclin-dependent kinases (CDKs), so inhibiting cell proliferation. When phosphorylated (P^5), $p27^{Kip1}$ might move into the cytoplasm, where, as shown by Besson *et al.*¹, it binds through its carboxy terminus (C) to RhoA and interferes with RhoA activation by guanine-nucleotide-exchange factors (GEFs). RhoA, Cdc42 and Rac regulate the cytoskeletal changes required for cell migration. Cdc42 and Rac work mainly at the front of polarized cells, regulating actin-driven protrusion and the formation of new adhesions required for forward movement. RhoA, through the ROCK protein, works mainly at the rear, determining (among other processes) the turnover of adhesive sites known as focal adhesions and thereby rear retraction. By interfering with RhoA activation, $p27^{Kip1}$ inhibits or promotes cell migration, depending on the cell type.

Whether or not this property does indeed contribute to cancer progression remains to be seen.

Cell proliferation is controlled by the sequential activation of a group of enzymes called cyclin-dependent kinases (CDKs), which are switched on when bound to cyclin proteins. The $p27^{Kip1}$ protein inhibits certain cyclin-CDK complexes in the cell nucleus, so preventing proliferation. The relevance of this protein to the control of proliferation is illustrated by the increased body size of $p27^{Kip1}$ -deficient mice. Studies of such animals have also revealed that, in general, a decrease or lack of $p27^{Kip1}$ increases the formation of tumours that are induced by various different oncogenic events, and also increases tumour-associated death rates^{2,3}.

In humans, too, abnormally low concentrations of $p27^{Kip1}$ are often found in tumours⁴. These low concentrations are also associated with tumour aggressiveness and patient mortality — characteristics that are generally related to tumour-cell invasion and metastasis. This raises the question of whether $p27^{Kip1}$ affects these types of tumour-cell properties, as well as proliferation. Last year, McAllister *et al.*⁵ obtained results that suggest this might be true — results that are now supported by Besson and colleagues¹. In short, these groups have shown that, when found in the cellular cytoplasm, $p27^{Kip1}$ influences cell migration.

Besson *et al.* also provide an explanation for this effect on cell migration: they find that $p27^{Kip1}$ binds to, and regulates the activity of, Rho proteins (Fig. 1). Members of the Rho family, which include Cdc42, Rac and RhoA, act as molecular switches in signalling path-

ways that affect gene transcription, as well as those that control the filaments of actin proteins in the cell's internal skeleton⁶. In particular, Rho proteins regulate and coordinate the cytoskeletal 'remodelling' that underlies changes in cell adhesion and migration. Rho proteins are switched on when bound to guanosine triphosphate (GTP), and off when bound to guanosine diphosphate (GDP). When active, they can interact with many different effector proteins that elicit different downstream responses. Various other proteins control the on/off state; guanine-nucleotide-exchange factors (GEFs), for instance, activate Rho proteins by promoting the replacement of GDP with GTP. Besson *et al.* show that $p27^{Kip1}$ inhibits RhoA activation by interfering with its interaction with GEFs.

The authors further find that the motility of $p27^{Kip1}$ -deficient fibroblast cell types decreases markedly in comparison with that of wild-type cells. Consistent with the earlier study⁵, re-expression of either wild-type $p27^{Kip1}$ or a truncated mutant that cannot bind to cyclin-CDK complexes restores migration — showing that the protein's regulation of motility is independent of its effects on proliferation. Indeed, the region of $p27^{Kip1}$ that binds RhoA and is required for migration is different from the region that binds cyclin-CDKs (Fig. 1).

The authors also discover that non-migratory, $p27^{Kip1}$ -deficient cells contain more actin stress fibres and focal adhesions — cellular features that are indicative of high RhoA activity — than wild-type cells. This could be confirmed by testing RhoA activity in cells. Interestingly, inhibiting the enzyme

ROCK, a downstream effector of RhoA, restores the migration of $p27^{Kip1}$ -deficient cells in response to growth factors. Apparently, then, in fibroblasts a decrease in $p27^{Kip1}$ concentration increases RhoA activity, which in turn increases ROCK activity, thereby hampering migration. Dampening Rho signalling, as occurs with normal $p27^{Kip1}$ concentrations or ROCK inhibition, allows fibroblasts to migrate again.

For tumours to metastasize, cells must alter their connections to their neighbours and their substrate, and then migrate. So could the effects of $p27^{Kip1}$ on migration somehow influence tumour progression? This is not easy to say, because models *in vitro* do not always mirror events *in vivo* — and different cell types vary in any case. For instance, efficient migration requires a tightly balanced activation and deactivation of Cdc42, Rac and RhoA in both space and time⁶. Low RhoA activity might inhibit migration by preventing cells from achieving sufficiently strong adhesion, and hence traction, to move forward. But excessive RhoA activity can also inhibit migration, by gluing cells to their substrate too strongly. Increased Rac-mediated cell-to-cell interactions also prevent migration⁷.

Still more complexity has been observed in a three-dimensional matrix⁸ in which cells can move in either an amoeboid or an elongated mode. Amoeboid migration is dependent on RhoA and ROCK, whereas the elongated mode is associated with Rac-dependent protrusions that are rich in actin filaments⁹. The degree of cell migration therefore depends on both the conditions and the cell type. So it is not surprising that the effects of $p27^{Kip1}$ on migration are also cell-type-dependent. For instance, over-expression of $p27^{Kip1}$ stimulates the migration of liver-cancer cells and fibroblasts^{1,5} but inhibits the migration of some other cell types, such as endothelial cells¹⁰. Consequently, aberrant $p27^{Kip1}$ signalling might promote the invasiveness and malignancy of some tumours but inhibit that of others. In support of this, a loss of $p27^{Kip1}$ usually correlates with increased tumour aggressiveness and a poor clinical outcome⁴. But in some tumour types, high $p27^{Kip1}$ concentrations do so^{11,12}.

One source of cell-type variability might lie in the fact that $p27^{Kip1}$ can regulate Rho activity only when $p27^{Kip1}$ is localized in the cytoplasm^{1,5}. Unlike other tumour suppressors, $p27^{Kip1}$ activity is downregulated by increased degradation or by exclusion from the nucleus. If the protein's concentration remains high but it is excluded from the nucleus, it can no longer regulate the cell cycle (so leading to increased proliferation), yet can interfere with RhoA and thereby influence cell migration and tumour progression. If its concentration is lowered but it remains in the nucleus, however, it again

cannot dampen cell proliferation, but neither can it influence migration.

Another confounding factor is that Rho proteins control a wide range of signalling pathways, which, when deregulated, can result in various cellular abnormalities¹³. As well as influencing migration, these proteins might affect cell survival, growth and differentiation⁶, thereby influencing tumour development and progression⁷. You won't be surprised to hear, then, that it will be a challenge to determine whether p27^{Kip1} contributes to tumour progression *in vivo* via regulation of the cell cycle, or Rho-mediated signalling, or both. The answer awaits the generation of mice in which wild-type p27^{Kip1} is replaced by a mutant version that lacks the binding domain for either cyclin-CDK complexes or RhoA. All being

well, these studies will reveal the true role of p27^{Kip1} in cancer. ■

John G. Collard is at the Netherlands Cancer Institute, Division of Cell Biology, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands.
e-mail: j.collard@nki.nl

1. Besson, A., Gurian-West, M., Schmidt, A., Hall, A. & Roberts, J. M. *Genes Dev.* **18**, 862–876 (2004).
2. Philipp-Staheli, J., Payne, S. R. & Kemp, C. J. *Exp. Cell Res.* **264**, 148–168 (2001).
3. Philipp-Staheli, J. *et al.* *Cancer Cell* **1**, 355–368 (2002).
4. Slingerland, J. & Pagano, M. J. *Cell. Physiol.* **183**, 10–17 (2000).
5. McAllister, S. S., Becker-Hapak, M., Pintucci, G., Pagano, M. & Dowdy, S. F. *Mol. Cell. Biol.* **23**, 216–228 (2003).
6. Ridley, A. J. *et al.* *Science* **302**, 1704–1709 (2003).
7. Malliri, A. & Collard, J. G. *Curr. Opin. Cell Biol.* **15**, 583–589 (2003).
8. Friedl, P. & Wolf, K. *Nature Rev. Cancer* **3**, 362–374 (2003).
9. Sahai, E. & Marshall, C. J. *Nature Cell Biol.* **5**, 711–719 (2003).
10. Goukassian, D. *et al.* *FASEB J.* **15**, 1877–1885 (2001).
11. Kouvaraki, M. *et al.* *Cancer* **94**, 2454–2465 (2002).
12. Saez, A. *et al.* *Br. J. Cancer* **80**, 1427–1434 (1999).
13. Boettner, B. & van Aelst, L. *Gene* **286**, 155–174 (2002).

Animal behaviour

Fickle females?

Michael J. Ryan

The courtship of satin bowerbirds is a complicated business. Different parts of a male's display appeal to females of different ages, so age-biased variation might underlie the evolution of these displays.

A common complaint among courting men is that women do not know what they want. Virgil, for instance, observed that "A woman is always a fickle, unstable thing". How, then, can a male cope? One solution would be to assess the general preference in this sea of shifting female wants and desires, and bias courtship to the majority opinion. Another would be to appeal simultaneously to multiple preferences, to be something to everyone. This second solution, we are told by Coleman and colleagues on page 742 of this issue¹, is the one adopted by male satin bowerbirds (*Ptilonorhynchus violaceus*).

These males have a complex set of courtship rituals. They decorate their bowers with blue objects, and they also combine a suite of vocalizations and movements into a behavioural display directed at females. Another layer of complexity derives from the timing of these display components. In the first stage of courtship a male is not present at the bower but his decorations are. Thus, the only way for a female to evaluate the male is through this evidence. In the second stage, a female returns to a subset of the original bowers she sampled. During this stage, both bower decorations and the male, who intensely courts the female, are present (Fig. 1), and the female builds nests at several bowers. In the final stage, the female makes several visits to where she constructed nests and decides on a single mate.

Coleman and colleagues' experiments¹ were underpinned by an observation from

a study² by members of the same group. Younger females, it turned out, are more threatened by the physical male displays than are older females. Could variation in this female fear factor generate selection for complex displays in males? In experiments conducted over a two-year period, Coleman *et al.* addressed the question by augmenting some bowers with extra decorations (treatment bowers) while leaving others unmanipulated (control bowers). Using video recordings, they determined the effect of bower treatment on the frequency of visits by females of different ages — first year, second year and three-plus years — at each of the three stages of courtship.

Of the bowers sampled during the first stage of courtship, all females were more likely to return to treated rather than control bowers in the second stage (remember, males are not present during the first stage). The same was not true, however, when females moved from the second stage, when the male is present and subjects the females to intense displays, to the third stage. Whereas younger females were more likely to visit the experimental than control bowers in the third stage, augmented bower decorations had no such effect on the behaviour of older females.

During the final stage of courtship, females decide on a mate. Here again there is an age-biased effect of bower treatment. In both years of Coleman and colleagues' study, first-year females preferred to mate with males in more decorated bowers, but the treatment did not influence the mate choice

of the three-plus-year females. In year 1 of the study, the second-year females' choice of mate showed no effect of bower treatment, but they did in year 2, when even more decorations were added.

An age-biased effect is not restricted to female responses to bower decorations, but also applies to male display intensity. The more intense the male display during the second stage of courtship, the more likely were older females to visit a male in the third stage. But that did not apply to the two younger cohorts. Coleman *et al.*¹ conclude that selection generated by age-biased variation in female choice is responsible for the evolution of complex courtship in these bowerbirds. The fear that the male's display instils in younger females seems to drive the age-biased differences in courtship preference.

Clearly, the further elaboration of bower decorations to the degree used in the manipulations would be under stronger selection from younger rather than older females. In this study, however, it is not clear if older females would generate selection to maintain the current level of bower decorations. Perhaps there is a lower threshold for this component of the display in older than in younger females; if so, older females alone could generate the selection needed to maintain this complex display. The fact that second-year females were not influenced by augmented bower decorations in year 1 of the study, but were in year 2 when more decorations were used, suggests that thresholds for an effect might change continuously with age. This age class could also offer a target group for research to understand the relative value of decorations and male display intensity for females of different ages.

A larger question concerns the evolutionary pattern of bowerbird displays. The complexity of bowers, decorations and courtship behaviour varies greatly among different species of bowerbird^{3,4}. Can these broad macroevolutionary patterns among species be explained by the same microevolutionary process suggested by the findings of Coleman *et al.*¹?

Studies of fish⁵, frogs⁶, cockroaches⁷ and humans⁸ have shown that female mating preferences can change as a function of female age, size and reproductive status. Such variation in preferences can contribute to the maintenance of variation in male traits. Young female guppies, for example, prefer males with more orange pigmentation, whereas older females are agnostic about colour and seem to prefer courtship vigour⁵; larger female cricket frogs prefer lower-frequency mating calls, but smaller females prefer higher-frequency calls⁶. What is unusual in the case of satin bowerbirds is that the variation in female mating preferences more clearly contributes to the evolution of multi-component and



Figure 1 Performance art — a male bowerbird displays as a female assesses the show from the bower.

multi-modal male displays. There is much interest in the functional significance of complex displays⁹, especially those that involve different senses. We see here that one explanation is that different components of such signals might be aimed at different audiences. We also see, at least in bowerbirds, that females aren't fickle: they just change their mind.

Michael J. Ryan is in the Section of Integrative Biology, University of Texas, Austin, Texas 78712, USA.
e-mail: mryan@mail.utexas.edu

1. Coleman, S. W., Patricelli, G. L. & Borgia, G. *Nature* **428**, 742–745 (2004).
2. Patricelli, G. L., Uy, J. A. & Borgia, G. *Behav. Ecol.* (in the press).
3. Gilliard, T. E. *Birds of Paradise and Bower Birds* (Weidenfeld & Nicholson, London, 1969).
4. Borgia, G. *Sci. Am.* **254** (June), 92–100 (1986).
5. Kodric-Brown, A. & Nicoletto, P. *Am. Nat.* **157**, 316–323 (2001).
6. Ryan, M. J., Perrill, S. A. & Wilczynski, W. *Am. Nat.* **139**, 1370–1383 (1992).
7. Moore, P. J. & Moore, A. J. *Proc. Natl Acad. Sci. USA* **98**, 9171–9176 (2001).
8. Wedekind, C., Seebeck, T., Bettens, F. & Paepke, A. *Proc. R. Soc. Lond. B* **260**, 245–249 (1995).
9. Partan, S. & Marler, P. *Science* **283**, 1272–1273 (1999).

Meteorology

Testing time for El Niño

David Anderson

Analyses that largely exploit indirect data from the past 150 years show that El Niño and La Niña might be more predictable than was thought. The results presage the prospect of extended climate forecasts.

Some of the worst famines in history were those of around 1877, a period in which perhaps 40 million people died in India and China¹. This and later disasters inspired meteorologists to try to predict whether the monsoon rains would be plentiful or sparse. The paper by Chen *et al.* on page 733 of this issue² is in this tradition. The authors have tested a model of the behaviour of the ocean and atmosphere to 'predict' the occurrence of ENSO (El Niño/Southern Oscillation), a phenomenon rooted in the equatorial Pacific Ocean but which has widespread climatic consequences.

From the early monsoon research, an appreciation emerged of the Southern Oscillation and some of its near-global impacts; later on, the connection to the periodic warming in the central-east equatorial Pacific Ocean known as El Niño was clarified. The science of seasonal forecasting and El Niño prediction was born as a result of the famines of the late 1800s, although it was a century before forecasts based on models of the ocean and atmosphere were developed. The term ENSO encompasses the cold as well as the warm state of the equatorial Pacific — that is, it includes La Niña as well as El Niño.

A major requirement for seasonal forecasting is an ability to predict ENSO events, which continue to have the potential to cause widespread disruption. For example, 1997–99 saw one of the largest El Niños in the instrumental record (Fig. 1, overleaf), followed by La Niña. Damage was estimated to exceed US\$20 billion, although beneficial effects in certain regions possibly offset this figure. If ENSO events can be predicted, steps can be taken to mitigate the losses. So successful forecasting of ENSO variability would be of great practical benefit. But such forecasting is immensely challenging, one difficulty being that the number of past events on which models can be tested is quite small owing to the lack of data. Chen *et al.*², however, have been able to test their model on events stretching back to before the 1877 El Niño, thus covering a much longer period than is normally considered.

From their results, the authors argue that a large ENSO event might be predictable two years in advance, much longer than hitherto expected. For example, their Fig. 3 (page 734) shows forecasts of six of the largest El Niños since 1856, including that of 1997–98. Even forecasts begun two years before the peak of the latter at the end of 1997 are reasonably good; and those from October 1996 capture the growth, maturity and decay in 1998 very well. These results are all the more impressive given that an earlier version of this model failed totally during the 'real-time' forecasting of the 1997–98 event. Many other models also performed poorly for this period³, although there was some success at the shorter range of a few months⁴.

So why are the new forecasts so much better than before? A major source of difficulty in climate prediction is model error, possibly the single biggest problem with physically based models. One way to deal with the issue is to improve the models, but this is a long, hard process; another is to accept that a model has error, but to try to mitigate the effects. That is what Chen *et al.* have done. To allow for model errors, corrections are made to the ocean state and to the sea surface temperature (SST) before its effects are 'passed' to the atmosphere; and corrections are made to the atmospheric wind before it is used to drive the ocean. ENSO forecasting requires knowledge of the ocean initial conditions and, to a small extent, those of the atmosphere. Through their strategy for correcting model error, Chen *et al.* also produce better initial conditions.

The 'training period' for deriving the model corrections is 1980–2000. As this includes both the 1997–98 event and the other big El Niño of 1982–83, there could be some artificial 'skill' in the predictions of this event. In general, the more choices and tunable parameters in a forecast methodology, the greater the risk of overestimating the forecast skill. Nonetheless, as the number of



Figure 1 El Niño makes itself known in September 1997. The haze in this picture of a boat on the Sarawak River, Borneo, is not an early morning mist but the smoke pollution from widespread fires that affected Malaysia and Indonesia. Because of El Niño, rainfall to cleanse the air of pollution was much reduced in this region, having been displaced far to the east.

past cases grows, the risk of this being a serious problem becomes less likely.

Increasing the number of cases on which to test a model's predictive capabilities is not straightforward, as there are no direct observations of ocean initial conditions before 1990. Considerable information about the ocean state can be gleaned from the winds near the Equator, but unfortunately, these were not adequately measured before the 1970s. The large-scale, slowly changing wind field can be derived from SSTs, however, and these are better known. Indeed, reconstructions go back to the middle of the nineteenth century and so ocean initial conditions can be indirectly generated. One might expect that the ocean initial conditions would be less well determined the further back in time one went, but Chen and colleagues' six-month forecasts for the 1877 El Niño are remarkably good, a result that would surely have delighted the forefathers of seasonal prediction.

For those involved in studies of ENSO evolution and predictability, much debate has centred on the importance of an energetic feature of the tropical atmosphere, the Madden–Julian Oscillation (MJO)^{3–8}; it has a timescale of the order of 50 days and can be associated with strong westerly winds near the Equator. In February 1997 there was a massive MJO in the west Pacific, which excited a big oceanic response that travelled along

the Equator all the way to the east Pacific and was commonly seen as the reason for the large 1997 El Niño. Chen and colleagues' model represents only the slow variations in the atmosphere and cannot represent MJO

variability. Nonetheless, forecasts from well before February 1997 show good predictive skill. From this and other results, they argue that MJO activity isn't as big a limitation on prediction as previously thought.

Overall, the authors give an optimistic outlook for ENSO forecasting, at least for large events: predicting smaller ones remains much trickier. Most operational models make predictions for six to nine months ahead of an ENSO onset. If the new results are confirmed, that range should be increased to at least a year. Rivaling or beating these forecasts is a challenge for more sophisticated global general-circulation models, such as those described in ref. 9. Another challenge, of course, concerns Chen and colleagues' model — will its success in predicting the future match that of its performance in predicting the past? ■

David Anderson is at the European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, UK.

e-mail: david.anderson@ecmwf.int

1. Kininmonth, W. *The 1997–98 El Niño Event: A Scientific and Technical Retrospective* (World Meteorological Organization, Geneva, 1999).
2. Chen, D., Cane, M. A., Kaplan, A., Zebiak, S. E. & Huang, D. *Nature* **428**, 733–736 (2004).
3. Landsea, C. W. & Knaff, J. A. *Bull. Am. Meteorol. Soc.* **81**, 2107–2119 (2000).
4. Stockdale, T. N., Anderson, D., Alves, J. & Balmaseda, M. *Nature* **392**, 370–373 (1998).
5. McPhaden, M. J. *Science* **283**, 950–954 (1999).
6. van Oldenborgh, G. J. *Mon. Weather Rev.* **128**, 2601–2607 (2000).
7. Moore, A. M. & Kleeman, R. Q. *J. R. Meteorol. Soc.* **122**, 1405–1446 (1996).
8. www.ecmwf.int/publications/library/do/references/list/17124
9. Vialard, J., Vitart, F., Balmaseda, M., Stockdale, T. & Anderson, D. ECMWF Tech. Memo 417 (2003); www.ecmwf.int/publications/library/ecpublications/_pdf/tm417.pdf

Mars

Blueberry fields for ever

Jeffrey M. Moore

The Mars saga continues. The latest finds — wide areas covered in balls of haematite, or 'blueberries', and large sulphate deposits in rocks — enable us to draw in more details of the planet's past climate.

We have known for thirty years that channels and valleys were carved into the martian landscape long ago by a fluid that was probably water. But unambiguous evidence for aqueous deposits has been frustratingly lacking. In 2000, images from the Mars Orbiter Camera, on board the Mars Global Surveyor, showed extensive rock outcrops with ledges and scarps, which were interpreted as being sedimentary layers that had been deposited under water¹. However, the failure of spectroscopic observations from orbiting instruments to reveal abundant water-modified minerals raised questions about the ubiquity and persistence of liquid water on the surface^{2,3}.

In the hope of building a clearer picture of water on Mars, NASA dispatched two rovers, Opportunity and Spirit, last summer to land at sites that were thought most likely to host accessible aqueous deposits. But before the rovers landed in January this year, the Mars Orbiter Camera provided the first incontrovertible evidence for the fluvial origin of a martian layered deposit: an exhumed river delta, resplendent with meanders and a recognizable upstream branching drainage basin^{4,5}.

Then, last month, came a flood of new evidence from the Opportunity rover and the European Space Agency's Mars Express orbiter. At briefings at NASA headquarters

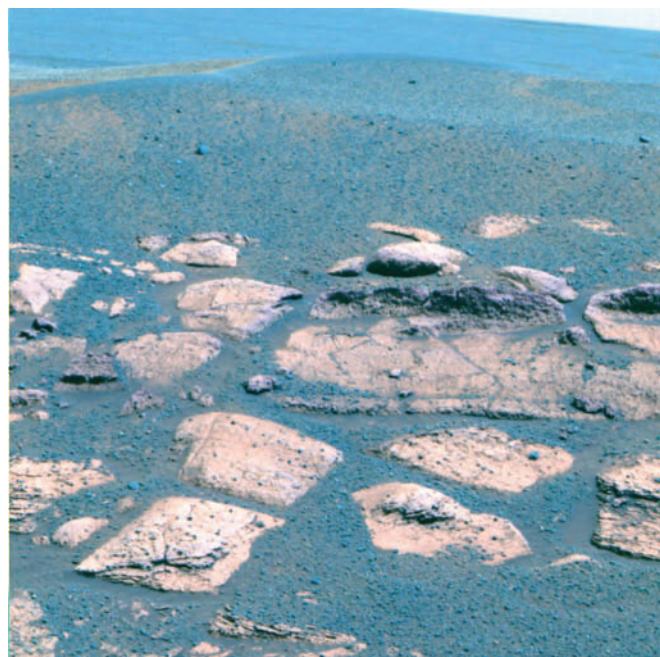


Figure 1 Blueberry way. Small spheres of haematite are strewn across the martian surface in this false-colour composite image from NASA's Opportunity rover. The outcrops in the foreground are the deposits of an ancient salty sea.

Mars sustained a thick CO_2 atmosphere, and supported liquid water, as long as the CO_2 was precluded from forming carbonates by SO_2 in the air and water. Sulphur dioxide very readily combines with cations to form sulphates, which seem now to make up the lion's share of Mars' subaqueous deposits. Once the abundance of SO_2 dropped below the critical level to suppress carbonate formation, the atmosphere would have rapidly collapsed to near its present size, leaving carbonates very little time to form as layered marine deposits.

The data from the Thermal Emission Spectrometer also hint that the sulphates formed before much of the present-day surface of the ancient southern highlands was carved out. The specular haematite deposits at Terra Meridiani probably formed at a temperature of about 100°C , converted from goethite^{10,11}. This conversion temperature implies that the blueberries on the lake beds of Terra Meridiani were once buried to a depth of more than a kilometre. There they might have lain for aeons, before the rock layers above were eroded away, revealing a field of blueberries that are too large to be transported or destroyed by the wind. The blueberry fields may have existed on the martian surface for millions of years.

The Mars of ancient times, of salty seas, does not readily advertise itself in the landscape that can be studied from orbit — a landscape that was probably shaped during the death throes of a warmer, wetter planet. We see the desiccated, elderly carcass and try to imagine the vigorous youth. The situation is akin to having topographic and spectral maps of the Grand Canyon on Earth, from which a good geologist could reasonably infer up to ten million years of fluvial erosion into flat-lying layered sediments. But the full history recorded in those sedimentary layers — the billion-year record of transgression and regression, sand ergs and dinosaurs — would be unfathomable¹². The rover results announced last month make it abundantly clear that our forensic interests in ancient Mars will require heavy investment in surface operations, and, ultimately, geologists toiling in the martian fields. ■

Jeffrey M. Moore is at the NASA Ames Research Center, MS 245-3, Moffett Field, California 94035, USA.

e-mail: jeff.moore@nasa.gov

1. Malin, M. C. & Edgett, K. S. *Science* **290**, 1927–1937 (2000).
2. Christensen, P. R. *et al.* *Science* **300**, 2056–2061 (2003).
3. Hoefen, T. M. *et al.* *Science* **302**, 627–630 (2003).
4. Malin, M. C. & Edgett, K. S. *Science* **302**, 1931–1934 (2003).
5. Moore, J. M., Howard, A. D., Dietrich, W. E. & Schenk, P. M. *Geophys. Res. Lett.* **30**, 2292–2295 (2003).
6. Pollack, J. B. *Icarus* **91**, 173–198 (1991).
7. Kahn, R. *Icarus* **62**, 175–190 (1985).
8. Clark, B. C. & Baird, A. K. *Geophys. Res. Lett.* **6**, 811–814 (1979).
9. Bandfield, J. L., Glotch, T. D. & Christensen, P. R. *Science* **301**, 1084–1087 (2003).
10. Christensen, P. R. *et al.* *J. Geophys. Res.* **105**, 9623–9642 (2000).
11. Catling, D. C. & Moore, J. M. *Icarus* **165**, 277–300 (2003).
12. Edgett, K. S. & Malin, M. C. *Lunar Planet. Sci.* **35**, 1188 (2004).

and at the Lunar and Planetary Science Conference*, Steve Squyres, principal investigator of the rover programme, announced that Opportunity was staring at an outcrop that had not only once been drenched in water, but that also bore the ripples and festoons of an ancient salty sea. The rover had quite improbably landed inside a rock-ledge-rimmed impact crater in the Terra Meridiani, a huge plain dominated by differentially eroded, flat-lying, layered sediments akin to Earth's Colorado Plateau. This area had been chosen for the landing as it was known to be rich in the mineral haematite, which might have formed in the presence of water. In perhaps some of the most striking images yet received from Opportunity, the haematite is manifest as berry-sized concretions, dubbed 'blueberries'. The blueberries litter the landing site to the horizon (Fig. 1), covering an area at least the size of my home state of Oklahoma.

The rocks at this site seem to be composed of up to 40% evaporite salts and minerals. Evaporites are precipitated rocks, potentially the best possible material for preserving not only microfossils (should they exist) but also trace chemical clues of early environmental and prebiotic evolution. But these are not just any evaporite minerals — they are sulphates, and not, interestingly, carbonates.

It had long been assumed that carbonates would be the most common, and stratigraphically uppermost, aqueous minerals on Mars⁶. The literature abounds with tales of the planet's supposed once-clement environment being destroyed by the removal of the greenhouse-supporting gas CO_2 from the atmosphere to make carbonates. (The

fact that the present martian atmospheric surface pressure sits at water's triple point is usually taken as evidence that this was so⁷.) But if sulphates, rather than carbonates, are the dominant evaporite, this thinking about the decline of the early martian climate might be all wrong.

The first results from OMEGA, the near-infrared spectrometer on board Mars Express, indicate that the sulphate deposit recorded by Opportunity is not a fluke. As reported at the conference, preliminary analysis of the high-resolution OMEGA data, covering much less than one per cent of the martian surface, has revealed a whopping sulphate signature in a light-toned, layered deposit on the floor of Valles Marineris, a system of canyons south of the martian equator (Y. Langevin, Univ. Paris-Sud).

If sulphates do indeed dominate the lacustrine record, it raises the possibility that the climate on the young planet was propped up by sulphur dioxide. Had the early martian waters contained even 0.1% SO_2 , dilute sulphuric acid would have formed, preventing the precipitation of carbonates⁸. Carbon dioxide would still have been the principal greenhouse gas in the atmosphere, but the vigorous belching of SO_2 from volcanoes might have kept the CO_2 from being taken up by rocks. As long as the SO_2 lasted, the CO_2 atmosphere could not collapse.

But Mars did eventually take up its atmospheric CO_2 . Data from the Thermal Emission Spectrometer on the Mars Global Surveyor show that modern martian dust, derived from weathered rock and other sources, contains magnesite (MgCO_3) at the level of between 2% and 5% by weight⁹; several bars' worth of CO_2 could be sequestered in this material (assuming that it infiltrated to a depth of 1–3 km). So it could be that

*35th Lunar and Planetary Science Conference, League City, Texas, USA, 15–19 March 2004.

Molecular biology

Yeast put through the changes with RNA

Genes Dev. 18, 794–804 (2004)

The fission yeast *Schizosaccharomyces pombe* exists in two mating types — or ‘sexes’ — known as P and M. Both types can divide to yield daughter cells of the opposite sex, but what triggers the ‘switching’ process has been unclear. Sonya Vengrova and Jacob Z. Dalggaard show that RNA is involved.

The yeast’s genome has a region called *mat1*, which contains one of two pairs of genes that determine mating type — one pair is found in the *mat1* region in P cells, the other in M cells. The ‘recombination’ event that underlies switching shuffles one set of genes or the other into the *mat1* region to be expressed, but quite how was not known. Vengrova and Dalggaard find that, during DNA replication, one or two nucleotides of RNA become inserted into the *mat1* region at a precise position. These ‘imprints’ are silent for one cell cycle, but during the next round they act as a block, halting DNA replication. This sparks off recombination and leads to mating-type switching.

Helen R. Pilcher

Meteorology

Dusty satellite data

Geophys. Res. Lett. doi:10.1029/2003GL019338 (2004)

The dust storms that blow across the Sahara Desert influence climate by absorbing or reflecting sunlight and altering the composition of clouds. Satellite data now show that existing analyses, used for various applications, may underestimate the speed at which dust plumes race away from one of the region’s most important birthplaces of these storms.

Ilan Koren and Yoram J. Kaufman studied data obtained by NASA’s Terra and Aqua satellites as they passed over the Bodele depression, northeast of Lake Chad. Here, a gap between the Tibesti and Ennedi mountains to the northeast forces winds across the depression, kicking up dust storms as they pass. As Aqua passes over the depression around three hours after Terra, the authors were able to follow the dust’s progress.

After examining 15 storms between January and April 2003, Koren and Kaufman calculate that the dust clouds are blown along at around 13 m s^{-1} — about double the speed calculated from previous ground-based measurements. What’s more, the authors estimate that a minimum wind speed of some 10 m s^{-1} is needed to kick a dust storm into action. When the dust reaches Africa’s west coast, however, where meteorological observatories are more common, the satellite and

ground-based data are in much better agreement.

Michael Hopkin

Ecology

Fungal stowaways

Mycol. Res. (in the press)

The US Army may have unwittingly killed hundreds of pine trees on a presidential Italian estate. Genetic analysis suggests that the trees were infected with an American fungus, perhaps imported by US troops during the Second World War.

Paolo Gonthier *et al.* analysed DNA samples taken from fungi at the base of seven dead trees. They found that the fungi were an eastern North American form of *Heterobasidion annosum*, a pest that decays the root system of pine trees. Four sequenced genes bore no relation to the European version of the fungus, matching instead the DNA of the North American species.

The researchers speculate that the fungus was transported to Europe by US troops in 1944, when they set up camp on the presidential estate of Castelporziano near Rome, which has been closed to the public for centuries. The pathogen might have stowed away in transport crates or other military equipment made from wood from infected North American trees.

Helen R. Pilcher

Physics

Liquid waves refract back

Phys. Rev. E 69, 030201 (2004)

Light waves refract when they travel from one material into another, bending towards — but never crossing — the normal (an imaginary line perpendicular to the interface between the two materials). In 2000, however, it was discovered that, in specially designed materials, electromagnetic waves can experience negative refraction, with the refracted ray being bent past the normal.

Xinhua Hu *et al.* have conducted a simple experiment to prove that surface waves on a

liquid can also refract in this unusual way. They filled a transparent tank with low-viscosity liquid and upright copper cylinders, then generated surface waves that travelled over the cylinders. As the wave frequency is increased, projected images of the waves show their focal point to be moving further away from the lens, and above about 5.8 Hz there is a transition from normal to negative refraction (pictured). The cylinders act as a ‘superlens’, bending the spreading water waves back towards a focal point on the opposite side of the barrier to the source. This matches theoretical predictions, and neatly mirrors the phenomenon of negative refraction seen for electromagnetic waves.

Mark Peplow

Nanoengineering

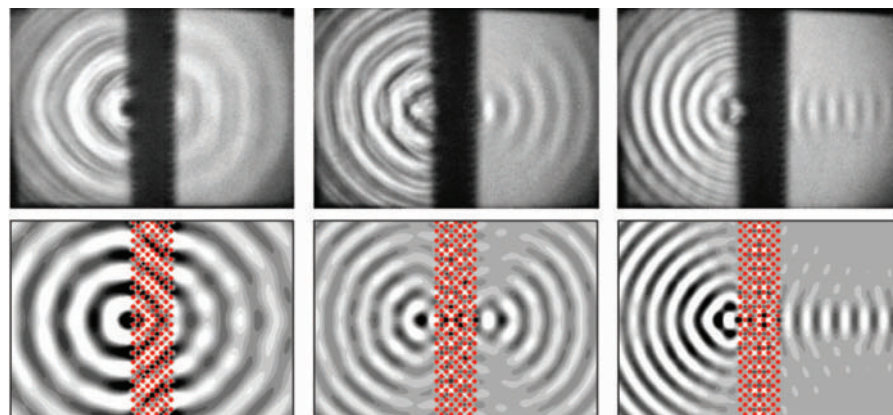
Much ado about nothing

Appl. Phys. Lett. 84, 2644–2645 (2004)

‘Nothing’ is a relative concept, but the emptiness of the little vacuum pocket created by Wei-Qiang Han and A. Zettl could hardly be more absolute. The so-called ultrahigh vacuum of conventional technology typically achieves pressures of less than 10^{-9} torr, but Han and Zettl have made a cavity in which they estimate the pressure to be around 10^{-84} torr. In comparison, outer space — at around 10^{-12} to 10^{-13} torr — is positively crowded.

The authors made their nanoscale cavity by using the electron beam of a transmission electron microscope to cleave a nanocrystal of potassium chloride that plugs the internal channel of a boron nitride nanotube. The nanocrystal has a width of 1.59 nm — six atomic layers — and fits almost perfectly within a nanotube of 2.3 nm inner diameter. Yet at room temperature the crystal slides spontaneously along the tube in a random walk. So when it is cleaved in two, the two fragments can move apart to create a cavity between them: an empty space preserved by an atomically tight seal at either end.

Phillip Ball



Waves on a liquid surface are refracted by a slab of copper cylinders (dark/red band). In experiment (top) and simulation (bottom), there is a transition from normal refraction to negative refraction as the frequency of the waves increases (left to right).

Sex differences in learning in chimpanzees

Young females quickly learn how to fish for termites — but young males prefer to play.

The wild chimpanzees in Gombe National Park, Tanzania, fish for termites with flexible tools that they make out of vegetation, inserting them into the termite mound and then extracting and eating the termites that cling to the tool¹. Tools may be used in different ways by different chimpanzee communities according to the local chimpanzee culture². Here we describe the results of a four-year longitudinal field study in which we investigated how this cultural behaviour is learned by the community's offspring. We find that there are distinct sex-based differences, akin to those found in human children, in the way in which young chimpanzees develop their termite-fishing skills.

Chimpanzees use tools for more purposes than any other non-human species³. The cultural variation in tool-use repertoires among chimpanzee communities may be attributable to individuals socially learning from other members of their community². We investigated this process in wild chimpanzees in Gombe National Park by videotaping 14 animals, who were all under 11 years old, and their mothers during termite-fishing sessions (for methods, see supplementary information).

We knew the ages of six individuals at the point when they first successfully obtained termites with a tool. In these individuals, we found that female offspring termite-fished when they were an average of 27 months younger than male offspring (for females: $n=3$, mean of 31 months, s.d. of 4 months; for males: $n=3$, mean of 58 months, s.d. of 6 months).

The increase over one-year age categories in the proportion of time spent attempting to termite-fish while at a termite mound, whether successful or not, closely resembles a logistic growth curve for both sexes (Fig. 1a). However, females termite-fished significantly more, and at an earlier age, than males in age classes from 0.5 to 5.5 years (general linear model⁴ (GLM): $n=25$, adjusted $r^2=0.82$; for age, sex, age $\times$ sex, $F_{10,14}=12.06$, $P<0.0001$). In young chimpanzees that had already acquired the skill (aged 3.5 to 10.5 years), females were also more proficient, as measured by the number of termites gathered per dip (GLM: $n=12$, adjusted $r^2=0.62$; age, $F_{1,9}=15.23$, $P=0.004$; sex, $F_{1,9}=10.94$, $P=0.009$) after adjusting for age.

Videotape recordings (E.V.L.) were used to categorize the length of the tool used by mothers and their offspring in terms of a chimpanzee's fist measure (see supplementary information): a tool inserted to less

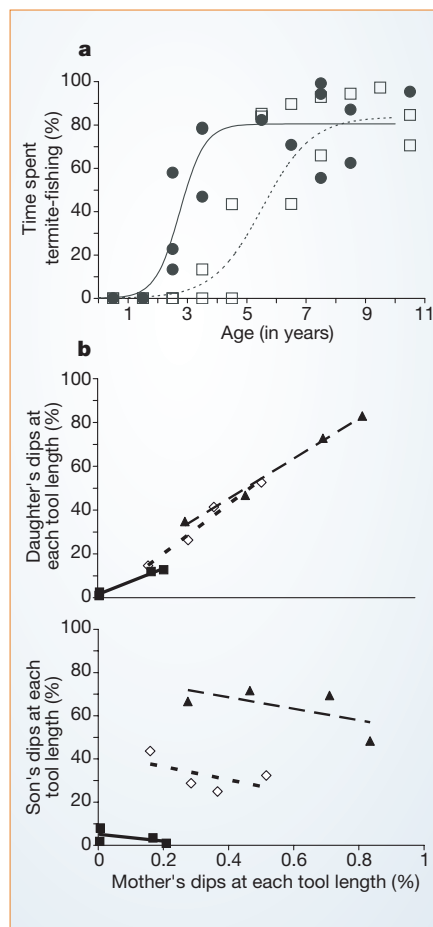


Figure 1 Sex differences in wild chimpanzees learning termite-fishing. **a**, Percentage of the time at the termite mound spent termite-fishing by young chimpanzees aged 0.5 to 10.5 years. Data are from 8 males (squares) and 6 females (circles) sampled over several years; fitted curves (dashed and full lines, respectively) represent the logistic prediction. **b**, Daughters' techniques (top) correlate strongly with their mothers' (for short tool insertion length (triangles), $P=0.005$, $r^2=0.99$; for medium tool insertion length (diamonds), $P=0.013$, $r^2=0.97$; for long tool insertion length (filled squares), $P=0.009$, $r^2=0.98$); sons' techniques (bottom) do not correlate with their mothers' (short, $P=0.378$, $r^2=0.39$; medium, $P=0.439$, $r^2=0.32$; long, $P=0.461$, $r^2=0.29$). Points represent an offspring's percentage of dippers at a particular tool insertion length plotted against his or her mother's percentage of dippers in the same length category. Regression lines are shown for each of the three categories of tool insertion length.

than 3 fists' length was categorized as short; 3–5 fists as medium; and more than 5 fists as long. Each mother showed a distinct pattern of apportioning her dippers among the different length categories. Daughters' termite-fishing techniques resembled their mothers with regard to tool insertion length, whereas the sons' did not.

For every tool insertion-length category, the proportion of a daughter's termite-

fishing dippers in that category correlated closely with her mother's (Fig. 1b, top); each son's percentage of dippers in a particular category did not correlate with his mother's (Fig. 1b, bottom). (GLMs for sex by mother's tool insertion-length category: short, $P=0.009$; medium, $P=0.01$; long, $P=0.02$.)

There were no significant differences between the sexes in the frequency of social interaction with the mothers, and mothers did not show any difference in tolerance towards male or female offspring. Because active demonstration of nut-cracking by a chimpanzee mother in the Tai forest has been described⁵, we looked for evidence of such behaviour in mothers at Gombe. We saw no cases of active teaching, by mothers or any other individuals, which would have been indicated, for example, by the offering of tools or modification of offspring behaviour.

Young females spent more time than young males watching their mothers fish for termites (generalized additive model⁶ (GAM): sex and age $\times$ sex interaction both give $P<0.0001$), whereas males spent significantly more time playing at the termite mound (GAM: sex and age $\times$ sex interaction both give $P<0.0001$).

Although our results are necessarily based on small numbers of chimpanzees, we used non-parametric models when standard model assumptions were violated, and tested for effects due to repeated measures where appropriate (see supplementary information). The sex differences in termite-fishing behaviours were consistent and strikingly apparent.

Our findings indicate that female chimpanzees start to fish for termites at a younger age than males; they are more proficient than males once they have acquired the skill; and they each use a technique similar to their mother's, although males do not. To our knowledge, this is the first systematic evidence of a difference between the sexes in the learning or imitation of a tool-use technique in wild chimpanzees. A similar disparity in the ability of young males and females to learn skills has been demonstrated in human children^{7–11} and may be indicative of different learning processes. A sex-based learning difference may therefore date back at least to the last common ancestor of chimpanzees and humans.

Elizabeth V. Lonsdorf*, Lynn E. Eberly†, Anne E. Pusey*

*Department of Ecology, Evolution and Behavior, University of Minnesota, Saint Paul, Minnesota 55108, USA

e-mail: elonsdorf@lpzoo.org

†Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota 55455, USA

- Goodall, J. *Symp. Zool. Soc. Lond.* **10**, 39–48 (1963).
- Whiten, A. et al. *Nature* **399**, 682–685 (1999).
- McGrew, W. C. *Chimpanzee Material Culture: Implications for Human Evolution* (Cambridge Univ. Press, Cambridge, 1992).
- McCulloch, C. E. & Searle, S. R. *Generalized, Linear, and Mixed Models* (Wiley, New York, 2000).
- Boesch, C. *Anim. Behav.* **41**, 530–532 (1991).
- Hastie, T. J. & Tibshirani, R. J. *Generalized Additive Models* (Chapman & Hall, London, 1990).
- Blote, A. W. & Van Gool, H. J. *Hum. Mov. Stud.* **17**, 133–152 (1989).
- Blote, A. W. & Van Haasteren, R. J. *Hum. Mov. Stud.* **17**, 187–205 (1989).
- Warrick, P. & Naglieri, J. J. *J. Edu. Psychol.* **85**, 693–701 (1993).
- Vederhus, L. & Kreckling, S. *Intelligence* **23**, 33–43 (1996).
- Davis, P. L. & Rose, J. D. *Phys. Occup. Ther. Ped.* **20**, 19–42 (2000).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Network dynamics

Jamming is limited in scale-free systems

A large number of complex networks are scale-free^{1,2} — that is, they follow a power-law degree distribution. Here we propose that the emergence of many scale-free networks is tied to the efficiency of transport and flow processing across these structures. In particular, we show that for large networks on which flows are influenced or generated by gradients of a scalar distributed on the nodes, scale-free structures will ensure efficient processing, whereas structures that are not scale-free, such as random graphs³, will become congested.

Many transport processes are induced by the existence of local gradients of some entity such as chemical potential, temperature or concentration. Gradients will also naturally generate, or influence, flows in complex networks. For example, in the case of information networks, properties of the nodes (such as rate of processing and adequacy) will generate a bias (formulated as a gradient criterion) in the way information is transmitted from a node to its neighbours. Specific examples include the World Wide Web, distributed computing⁴ and social networks with competitive dynamics⁵.

A simple model of a transport process begins by assuming that there are N nodes whose connections are described by a fixed substrate network, S . Associated with each node i is a non-degenerate scalar, h_i , which describes the 'potential' of the node. Then the gradient network can be constructed as the collection of directed links pointing from each node to whichever of its near-neighbours on S or itself has the highest potential. It can be shown that all non-degenerate gradient networks are forests — that is, they have no loops (except for self-loops) and consist only of trees. Furthermore, if S is a simple random graph³, in

which each pair of nodes is linked with probability P , and the scalars h_i are independent identically distributed random variables, then the distribution of the number of links pointing to each node (the in-degree distribution) becomes (equation (1); derivation to be published elsewhere)

$$R(l) = \frac{1}{N} \sum_{n=0}^{N-1} \binom{N-1-n}{l} (1 - P(1-p)^n)^{N-1-n-l} (P(1-p)^n)^l$$

In the limit $N \rightarrow \infty$ and $P \rightarrow 0$, such that the average degree $z \equiv NP = \text{constant} \gg 1$, the exact degree distribution (equation (1)) becomes the power law $R(l) \approx l^{-1}$, with a finite-size cut-off at $l_c = z$ (Fig. 1a). In this limit, therefore, gradient networks are scale-free. This is surprising⁶ because the substrate S is not scale-free, and in the same limit it has a 'bell-curve' degree distribution. Alternatively, if the substrate network S is scale-free,

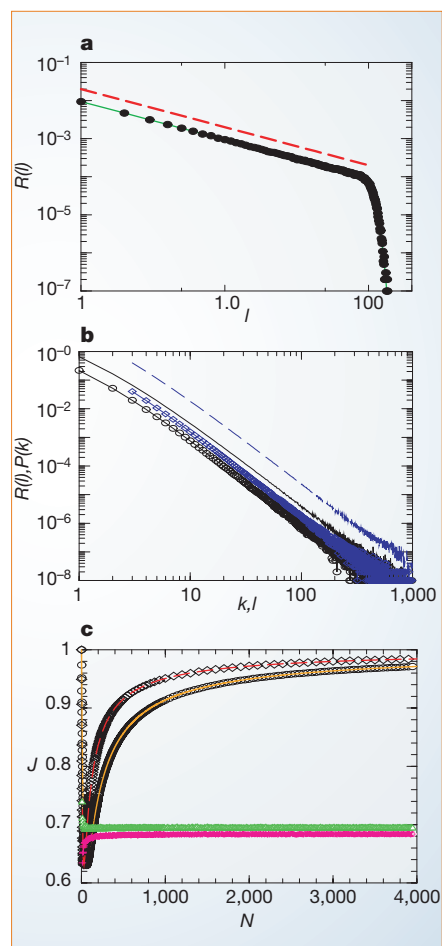


Figure 1 Degree distributions of gradient networks and jamming. **a**, Comparison between the exact formula (equation (1), green line) and numerical simulations (ellipses). Here $N=1,000$, $P=0.1$ ($z=100$); slope (dashed red line) is -1 . **b**, Degree distributions of the gradient network and the substrate, when the substrate is a Barabási–Albert scale-free graph with parameter m (for $m=1$, in black: $R(l)$, circles; $P(k)$, full line. For $m=3$, in purple: $R(l)$, diamonds; $P(k)$, dashed line; $N=10^5$). **c**, Jamming coefficient J for random graphs (circles, $P=0.05$; diamonds, $P=0.1$; orange line, $P=0.05$, exact expression; red line, $P=0.1$, exact expression) and scale-free networks (pink, $m=1$; green, $m=3$). Each data point is the average of 10^4 runs in **a** and 10^3 runs in **b**.

as in a Barabási–Albert network¹, then the associated gradient network is also scale-free (Fig. 1b) and is characterized by the same exponent.

For a network S , where the flow is processed at nodes in a finite time (such as in the routing of a packet), the quality of flow processing can be quantified by considering the jamming or congestion factor, which is defined as (equation (2))

$$J = 1 - \left\langle \frac{N_{\text{receive}}}{N_{\text{send}}} \right\rangle_{h_i \text{ network}} = R(0)$$

where N_{receive} is the number of nodes that receive gradient flow and N_{send} is the number of nodes that send it. Note that J is a queuing characteristic, rather than an actual throughput measure. Certainly, $0 \leq J \leq 1$, with $J=1$ corresponding to maximal congestion (vanishing number of receivers/processors) and $J=0$ to no congestion. For a random substrate network, the expression of J simply follows from equations (1) and (2) which, in the large network scaling limit, where P is constant and $N \rightarrow \infty$, becomes

$$J(N, P) = 1 - \frac{\ln N}{N \ln \left(\frac{1}{1-P} \right)} \left[1 + O\left(\frac{1}{N}\right) \right] \rightarrow 1$$

where $O(1/N)$ indicates corrections of order $1/N$.

Random networks therefore become maximally congested in that limit. For scale-free networks, however, the conclusion in the same limit is drastically different. In that case, J tends to a positive constant bounded away from unity — that is, scale-free networks are not prone to jamming. Figure 1c compares the congestion as a function of N for both random and scale-free substrate networks.

Zoltán Toroczkai*, Kevin E. Bassler†

*Center for Nonlinear Studies and Complex Systems Group, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA
e-mail: toro@lanl.gov

†Department of Physics, University of Houston, Houston, Texas 77204-5005, USA

- Barabási, A. L. & Albert, R. *Science* **286**, 509–512 (1999).
- Newman, M. E. J. *SIAM Rev.* **45**, 167–256 (2003).
- Bollobás, B. *Random Graphs* 2nd edn (Cambridge Univ. Press, Cambridge, 2001).
- Rabani, Y., Sinclair, A. & Wanka, R. *Proc. 39th Symp. Found. Comp. Sci.* 694–706 (1998).
- Anghel, M., Toroczkai, Z., Bassler, K. E. & Korniss, G. *Phys. Rev. Lett.* **92**, 058701 (2004).
- Lakhina, A., Byers, J. W., Crovella, M. & Xie, P. *Proc. IEEE Infocom* 03 (2003).

Competing financial interests: declared none.

brieF communications arising online

www.nature.com/bca

Pathology: Whales, sonar and decompression sickness

C. A. Piantadosi & E. D. Thalmann
(doi:10.1038/nature02527)

Reply: A. Fernández et al.
(doi:10.1038/nature02528)

Pathology

Whales, sonar and decompression sickness

Arising from: Jepson, P. D. *et al. Nature* **425**, 575–576 (2003).

We do not yet know why whales occasionally strand after sonar has been deployed nearby, but such information is important for both naval undersea activities and the protection of marine mammals. Jepson *et al.* suggest that a peculiar gas-forming disease afflicting some stranded cetaceans could be a type of decompression sickness (DCS) resulting from exposure to mid-range sonar¹. However, neither decompression theory nor observation support the existence of a naturally occurring DCS in whales that is characterized by encapsulated, gas-filled cavities in the liver. Although gas-bubble formation may be aggravated by acoustic energy, more rigorous investigation is needed before sonar can be firmly linked to bubble formation in whales.

On the basis of the available information, the DCS hypothesis of Jepson *et al.* contains two flaws. First, whales do not develop sufficient gas supersaturation in the tissues on ascent to cause extensive bubble formation in the liver. The gas available for supersaturation is limited to that present in the lungs at the onset of each held breath. During descent, the thorax is compressed², and the residual gas volume in the compliant lungs is forced, by Boyle's law contraction and alveolar collapse, into non-respiratory conducting airways, where it is sequestered from the circulation³. Not enough gas is taken up to produce bubbles, except possibly during multiple rapid dives to depths approaching that of the lung's closing volume³.

Once nitrogen uptake is blocked by lung collapse, the partial pressure of nitrogen in the blood actually decreases for the rest of the dive as nitrogen is distributed to tissues⁴. However, nitrogen accumulation in the liver, intestines and other visceral organs is limited by the diving response, which directs arterial blood away from these organs to the brain and heart⁵. On ascent, bubbles leaving supersaturated tissues must enter the venous blood and return to the lungs. These bubbles cannot pass into the liver (except in the unlikely event that they are of portal origin) unless they bypass the lung, which serves as a bubble trap. Jepson *et al.* do not explain why DCS, if it did occur in the whales they investigated, should affect the liver disproportionately.

Second, large gas-filled cavities in the liver, many encapsulated in dense fibrous tissue, are inconsistent with the pathology of DCS in humans and other mammals in which the bones, joints, lungs and central nervous system are primarily affected. The liver is rarely involved, and never to the extent described by Jepson *et al.* in the

cetaceans that they investigated, even when the body's gas burden is very large. Bubbles have been observed in hepatic sinusoids and in the portal vein, but large encapsulated bubbles have not been reported⁶. Liver lesions are equally unexpected in 'recurrent' DCS, in which chronic lesions are found only in the long bones and the central nervous system. It is unlikely that fibrotic hepatic lesions of the type described by Jepson *et al.*, which would require days or weeks to develop, could be caused by a brief exposure to sonar.

We agree with Jepson *et al.* that further investigation is needed, including an analysis of the composition of the gas in the bubbles. But we believe that identifying the cetacean gas disease with DCS is premature because its pathology not only differs from that underlying the syndrome in other mammals, but it also cannot be explained by any physiological mechanism related to diving.

Claude A. Piantadosi, Edward D. Thalmann

Center for Hyperbaric Medicine and Environmental Physiology, Duke University Medical Center, Durham, North Carolina 27710, USA

e-mail: piant001@mc.duke.edu

doi:10.1038/nature02527

1. Jepson, P. D. *et al. Nature* **425**, 575–576 (2003).
2. Williams, T. M. *et al. Science* **288**, 133–136 (2000).
3. Houser, D. S., Howard, R. & Ridgway, S. J. *Theor. Biol.* **213**, 183–195 (2001).
4. Falke, K. J. *et al. Science* **229**, 556–558 (1985).
5. Kooyman, G. L. & Ponganis, P. J. *Annu. Rev. Physiol.* **60**, 19–32 (1998).
6. Francis, T. J. R. & Mitchell, S. J. in *Bennett and Elliott's Physiology and Medicine of Diving* (eds Brubakk, A. O. & Neuman, T. S.) 5th edn 530–556 (Saunders, Philadelphia, 2003).

Jepson et al. reply — We did not, as Piantadosi and Thalmann suggest¹, present our findings as conclusive evidence of decompression sickness (DCS). We stated neither that DCS occurs naturally in cetaceans, nor that exposure to active sonar increases its occurrence². However, we restate that there is now a generally accepted link between some beaked-whale strandings and sonar use, and that lesions in some cetaceans demonstrate that *in vivo* bubble formation (embolus) can occur and persist.

Progressively increasing concentrations of nitrogen in cetacean tissues after repetitive diving have been studied empirically in bottlenose dolphins (*Tursiops truncatus*)³ and higher levels are predicted for northern bottlenose whales (*Hyperoodon ampullatus*) on the basis of their rate of descent or ascent and depth of diving⁴. Nitrogen supersaturation could be further increased by an accelerated rate of ascent, possibly to a critical point where bubbles form.

Even if naturally occurring levels of

nitrogen supersaturation in the tissues of diving cetaceans are normally insufficient to initiate bubble formation, a theoretical possibility remains that cetaceans with nitrogen-supersaturated tissues could experience bubble growth or formation as a result of intense acoustic exposure^{4,5}. There was a clear spatial and temporal link with active naval sonar exposure in the case of the beaked whales in the Canary Islands, as well as in previously reported beaked-whale strandings^{6,7}.

The lesions in the Canary Island beaked whales and in the UK cases (mainly dolphins) differed. The beaked whales had acute, systemic and widely disseminated lesions consistent with, although not diagnostic of, DCS⁸. The large hepatic cavities found exclusively in the UK cases are atypical of DCS in humans and experimental animals. For logistical reasons, the central nervous system was examined in only two UK cases and the bones were not examined in any. We cannot therefore confirm or refute the presence of lesions consistent with gas embolism in bone or the central nervous system. However, large numbers of gas bubbles were seen in portal veins and sinusoids in the livers from all UK cases examined microscopically, consistent with DCS in humans⁸.

As cetaceans differ from humans behaviourally (as obligate, repetitive breath-hold divers), physiologically (for example, in their diving reflex⁹ and anatomically (as in their retia mirabilia, large portal veins and diaphragmatic sphincters)^{10,11}), it may be too simplistic to assume that the distribution, severity and chronicity of lesions induced by gas emboli will be similar in both human divers and free-living cetaceans. Extensive sublethal bubble formation in human DCS is an acute medical emergency. Without medical intervention, a free-living cetacean suffering the same fate would continue diving for days, weeks or months afterwards unless death or stranding intervened.

Lesion pathogenesis in the stranded cetaceans² may ultimately be explained by bubble formation, possibly in response to either rapid decompression or acoustic exposure of nitrogen-supersaturated tissues⁵; however, it is not clear how marine mammals mitigate the accumulation of nitrogen gas while diving and defend themselves against nitrogen-bubble formation. These uncertainties do indeed argue for caution in interpreting the limited studies available.

A. Fernández*, M. Arbelo, R. Deaville, I. A. P. Patterson, P. Castro, J. R. Baker, E. Degollada, H. M. Ross, P. Herráez, A. M. Pocknell, E. Rodríguez, F. E. Howie, A. Espinosa, R. J. Reid, J. R. Jaber, V. Martin, A. A. Cunningham, P. D. Jepson

**Histology and Pathology Unit, Institute for Animal Health, Veterinary School, Montana Cardones–Arucas, University of Las Palmas de Gran Canaria, Gran Canaria, Spain*

e-mail: afernandez@dmor.ulpgc.es

doi:10.1038/nature02528

1. Piantadosi, C. A. & Thalmann, E. D. *Nature* doi:10.1038/nature02527 (2004).
2. Jepson, P. D. *et al. Nature* **425**, 575–576 (2003).
3. Ridgway, S. H. & Howard, R. *Science* **206**, 1182–1183 (1979).
4. Houser, D. S., Howard, R. & Ridgway, S. H. *J. Theor. Biol.* **213**, 183–195 (2001).
5. Crum, L. A. & Mao, Y. *J. Acoust. Soc. Am.* **99**, 2898–2907 (1996).
6. *US Department of Commerce and US Navy Joint Interim Report* http://www.nmfs.noaa.gov/prot_res/overview/Interim_Bahamas_Report.pdf (2001).
7. Frantz, A. *Nature* **392**, 29 (1998).
8. Francis, T. J. R. & Mitchell, S. J. in *Bennett and Elliott's Physiology and Medicine of Diving* (eds Brubakk, A. O. & Neuman, T. S.) 5th edn 530–556 (Saunders, Philadelphia, 2003).
9. Ridgway, S. H. in *Mammals of the Sea: Biology and Medicine* (ed. Ridgway, S. H.) 590–747 (Thomas, Springfield, Illinois, 1972).
10. Harrison, R. J. & Tomlinson, D. W. *Proc. Zool. Soc. Lond.* **126**, 205–233 (1956).
11. Simpson, J. G. & Gardener, M. B. in *Mammals of the Sea: Biology and Medicine* (ed. Ridgway, S. H.) 298–418 (Thomas, Springfield, Illinois, 1972).

Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks

Michael D. Shapiro^{1*}, Melissa E. Marks^{1*}, Catherine L. Peichel^{2*}, Benjamin K. Blackman¹, Kirsten S. Nereng¹, Bjarni Jónsson³, Dolph Schluter⁴ & David M. Kingsley¹

¹Department of Developmental Biology and HHMI, Stanford University School of Medicine, Stanford, California 94305-5329, USA

²Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA

³Institute of Freshwater Fisheries, Holar Agricultural College, IS-551 Saudarkrokur, Iceland

⁴Department of Zoology, University of British Columbia, 6270 University Blvd, Vancouver V6T 1Z4, Canada

* These authors contributed equally to this work

Hindlimb loss has evolved repeatedly in many different animals by means of molecular mechanisms that are still unknown. To determine the number and type of genetic changes underlying pelvic reduction in natural populations, we carried out genetic crosses between threespine stickleback fish with complete or missing pelvic structures. Genome-wide linkage mapping shows that pelvic reduction is controlled by one major and four minor chromosome regions. *Pitx1* maps to the major chromosome region controlling most of the variation in pelvic size. Pelvic-reduced fish show the same left-right asymmetry seen in *Pitx1* knockout mice, but do not show changes in *Pitx1* protein sequence. Instead, pelvic-reduced sticklebacks show site-specific regulatory changes in *Pitx1* expression, with reduced or absent expression in pelvic and caudal fin precursors. Regulatory mutations in major developmental control genes may provide a mechanism for generating rapid skeletal changes in natural populations, while preserving the essential roles of these genes in other processes.

Vertebrate limb structures exhibit extensive structural variation in animals adapted to different environments. In a phylogenetically diverse set of vertebrates—including lineages of reptiles, amphibians, marine mammals and fish—hindlimb structures are characteristically reduced or missing. The presence of rudimentary hindlimb structures in snakes and whales formed part of the early evidence indicating that these animals evolved from four-limbed ancestors through extensive modification of pre-existing skeletal structures^{1,2}. Despite longstanding interest in the mechanisms that control hindlimb reduction during vertebrate evolution, the detailed number, location and type of genetic changes that underlie this process are still unknown^{3–5}.

Marine threespine sticklebacks (*Gasterosteus aculeatus*) have a prominent pelvic skeleton made up of bilateral pelvic spines that articulate in trochlear joints with an underlying pelvic girdle. The girdle covers part of the ventral surface and extends up the lateral side of the fish in an ascending branch that articulates with dermal armour plates (Fig. 1f). Previous studies suggest that pelvic structures protect sticklebacks against gape-limited, soft-mouthed predators by presenting a lacerating defensive structure, which increases the effective diameter of the fish and resists compressive forces during predator manipulation and chewing^{6,7}. However, several freshwater stickleback populations have evolved complete or partial loss of the pelvic skeleton^{8–12}, perhaps in response to local absence of predatory fish, reduced levels of calcium ion availability, or predation by macroinvertebrates that typically capture sticklebacks by grasping the dorsal and pelvic spines^{7,13–17}. The young geological age of the postglacial lakes containing pelvic-reduced sticklebacks, and the rapid tempo of pelvic changes seen in a high-resolution series of fossil sticklebacks¹⁸, suggest that pelvic reduction can evolve in less than 10,000 generations in this system.

control pelvic reduction in natural populations, we crossed a marine female showing robust development of pelvic structures to a freshwater male from the Paxton Lake benthic population¹⁹. As with the majority of the Paxton benthic population, this fish had no pelvic spines or underlying pelvic structures. A total of 375 adult F₂ progeny from a single pair of F₁ hybrid parents, both with complete pelves, were killed, measured for extent of development of several different pelvic features (Fig. 1f), and genotyped with a large set of previously described microsatellite markers²⁰.

Scoring pelvic reduction as a qualitative trait (normal pelvic structures versus any form of size reduction, loss or asymmetry) revealed a near 3:1 mendelian ratio of unaffected to affected fish (289 and 86 progeny, respectively). Mapping of this qualitative trait revealed strong linkage to markers on the distal end of linkage group 7 (log likelihood ratio of linkage (LOD) = 36.9). The same major chromosome region also controlled a substantial portion (13.5–43.7%) of the variance in every quantitative measure of pelvic size in the cross, including lengths of the pelvic spines and pelvic girdles, and heights of the ascending branches of the pelvis (Table 1 and Fig. 1). Several chromosome regions with smaller effects (minor modifier quantitative trait loci (QTL)) were also detected in the cross, each controlling between 5.6% and 11.1% of the variance in specific pelvic measurements, and mapping to several different linkage groups (Fig. 1b–e). The remaining variance may be due to additional minor loci with phenotypic effects that are too small to detect in a cross of this size, or to environmental and epigenetic factors. Each of the QTL detected in the current cross mapped to regions distinct from a previously reported QTL that modified pelvic spine length in fish from Priest Lake, British Columbia²⁰.

At each chromosome region that influenced pelvic morphology in the cross between marine and Paxton benthic sticklebacks, the net effect of freshwater alleles from the Paxton grandparent was to decrease the average size of pelvic structures (Table 1). Even fish with one marine and one Paxton benthic allele near a given QTL

Genetic architecture of pelvic reduction

To examine the number and location of chromosome regions that

showed a significant change in the mean size of pelvic structures compared with fish with two marine alleles at that locus (see mean phenotypes as a function of heterozygosity or homozygosity for Paxton benthic alleles near QTL on linkage groups 1, 4, 7 or 10 in Table 1). Increasing the total number of Paxton benthic alleles at the minor QTL on linkage groups 1, 2 and 4 produced larger combined effects than each minor modifier QTL individually (Fig. 1g, h). The magnitude of the phenotypic effect of the modifier QTL varied in fish that inherited only marine, marine and benthic, or only benthic alleles at the major QTL on linkage group 7 (see Supplementary Information). Both additive and epistatic interactions between loci are therefore likely to influence pelvic morphology. Rare, favourable, new alleles with semi-additive effects are expected to spread more rapidly in an evolving population than purely recessive alleles, when the alleles begin at low frequency²¹. In addition, traits controlled by a relatively small number of genes should evolve more quickly than traits that require the separate origin and fixation of new mutations at many different loci. Our results suggest that major morphological changes in the vertebrate hindlimb skeleton may occur through relatively few chromosome regions in natural populations. The heterozygous effects of Paxton benthic alleles, and the existence of a single major chromosome region controlling up to half of the phenotypic variance, may help to explain the rapid tempo of pelvic reduction in the stickleback system.

Isolation and mapping of candidate genes

Several genes have been described that are expressed specifically in hindlimbs but not forelimbs, or are required for normal hindlimb development in traditional vertebrate model systems. These genes

include the transcription factors *Tbx4* (refs 22–24), *Pitx1* (refs 24–26) and *Pitx2* (ref. 27). To test whether variation in these genes may contribute to pelvic reduction in the Paxton benthic population, we isolated bacterial artificial chromosome (BAC) clones containing stickleback orthologues of each gene, sequenced portions of the clones to identify polymorphic genetic markers, and determined the segregation pattern of these markers in F₂ animals from the marine by Paxton benthic cross. The stickleback *Tbx4* and *Pitx2* markers mapped to linkage groups 1 and 4, respectively (Fig. 1b, d), clearly excluding them from the major chromosome region at the distal end of linkage group 7, which has the largest effect on pelvic reduction in the cross.

In contrast, a microsatellite marker located in the first intron of the *Pitx1* gene mapped to the distal end of linkage group 7, in a position indistinguishable from the major region controlling pelvic reduction (Fig. 1a). Addition of *Pitx1* to the linkage map substantially increased both the overall LOD score and the total variance explained by linkage group 7 for all the traits (Table 1), consistent with tight linkage between *Pitx1* and the pelvic reduction phenotype. A total of 70 F₂ progeny that showed complete bilateral absence of pelvic spines were identified from the mapping family and two additional families. Each of these fish was homozygous for Paxton benthic alleles at the *Pitx1* locus, showing that *Pitx1* does not recombine with the severe pelvic reduction phenotype in at least 140 chromosomes.

Left–right asymmetry in stickleback pelvic reduction

Pitx1 is a particularly intriguing candidate for the major locus controlling pelvic reduction in sticklebacks. During normal devel-

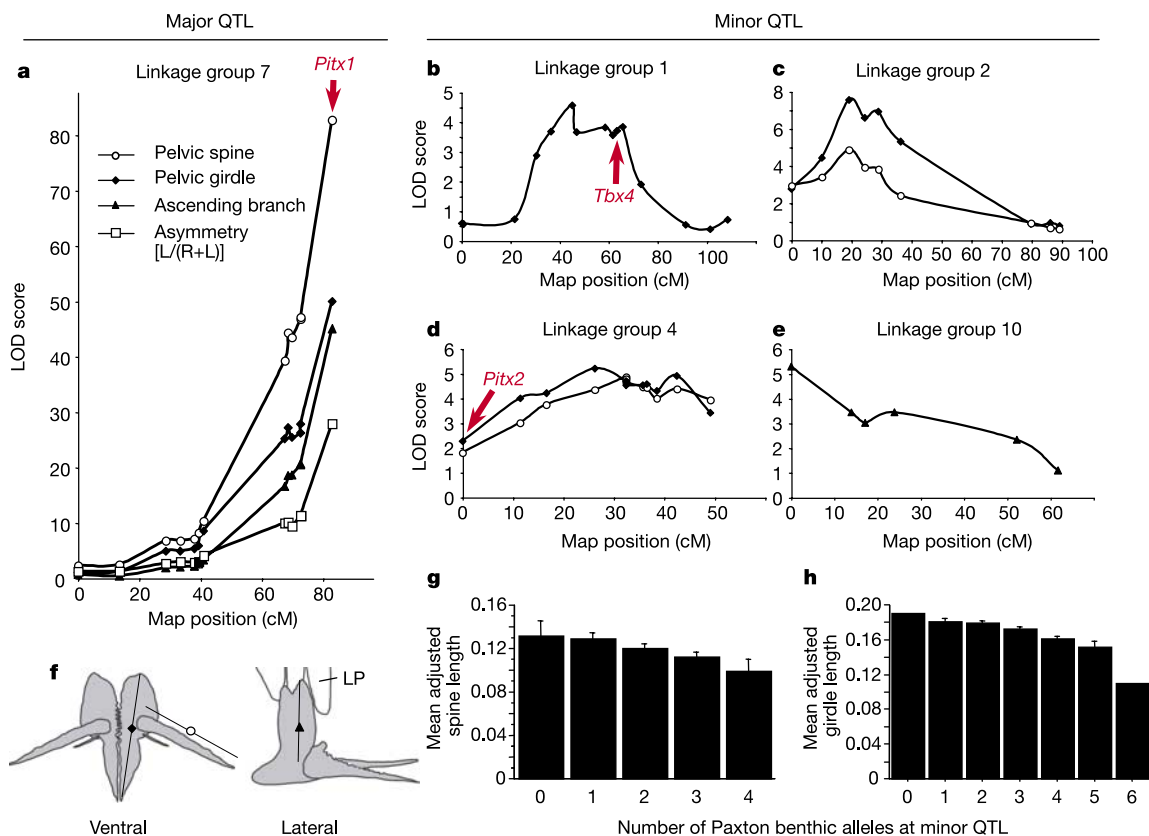


Figure 1 Genetic architecture of pelvic reduction in a cross between marine and Paxton benthic sticklebacks. **a–e**, LOD scores for different pelvic traits are graphed as a function of genetic distance (centimorgans; cM) along linkage groups²⁰ with significant QTL (LOD ≥ 4.5 (ref. 50)). (See Supplementary Information for markers.) **f**, Different quantitative measurements of pelvic structures. LP, lateral plates. **g, h**, Increasing

substitution of Paxton benthic alleles at minor loci affecting spine length (**g**, linkage group 2, 4) or pelvic girdle length (**h**, linkage group 1, 2, 4) leads to a progressive reduction in the mean phenotypic values for each trait (multiple-way analysis of variance: $P < 0.03$, $P < 0.0001$, respectively). Error bars indicate standard error.

Table 1 Location and magnitude of effect for major and minor pelvic reduction QTL

Trait	LG	Marker	Map position (cM)	LOD	PVE (%)	Phenotype means		
						MM	MB	BB
Complete versus reduced pelvis	7	<i>Pitx1</i>	82.9	72.6	NA	NA	NA	NA
	7	<i>Stn82</i>	72.6	36.9	NA	NA	NA	NA
Pelvic spine length	7	<i>Pitx1</i>	82.9	82.8	65.3	0.148 ± 0.002	0.132 ± 0.001†	0.068 ± 0.004‡§
	7	<i>Stn82</i>	72.6	45.5	43.7	0.144 ± 0.002	0.126 ± 0.002†	0.072 ± 0.005‡§
	2	<i>Stn21</i>	19.1	4.9	7.6	0.115 ± 0.011	0.131 ± 0.005	0.098 ± 0.014‡
	4	<i>Gac4174</i>	32.4	4.9	5.8	0.134 ± 0.005	0.118 ± 0.003*	0.108 ± 0.004†
Pelvic girdle length	7	<i>Pitx1</i>	82.9	50.0	46.8	0.185 ± 0.001	0.178 ± 0.001*	0.140 ± 0.003‡§
	7	<i>Stn82</i>	72.6	25.6	27.8	0.183 ± 0.003	0.174 ± 0.003*	0.144 ± 0.002‡§
	1	<i>Stn7</i>	45.1	4.6	5.6	0.178 ± 0.002	0.170 ± 0.002*	0.160 ± 0.004‡†
	2	<i>Stn21</i>	19.1	7.6	11.1	0.169 ± 0.005	0.177 ± 0.004	0.147 ± 0.008*‡
	4	<i>Gac4174</i>	32.4	4.7	5.6	0.178 ± 0.002	0.171 ± 0.002	0.159 ± 0.004‡†
Ascending branch height	7	<i>Pitx1</i>	82.9	45.1	44.5	0.108 ± 0.001	0.106 ± 0.001	0.077 ± 0.002‡§
	7	<i>Stn82</i>	72.6	19.8	22.2	0.107 ± 0.001	0.100 ± 0.002*	0.079 ± 0.003‡§
	10	<i>Stn119</i>	0	5.3	6.6	0.105 ± 0.001	0.099 ± 0.002*	0.091 ± 0.002‡†
Asymmetry	7	<i>Pitx1</i>	82.9	28.0	31.5	0.504 ± 0.006	0.502 ± 0.001	0.701 ± 0.029‡§
	7	<i>Stn82</i>	72.6	11.2	13.5	0.506 ± 0.006	0.525 ± 0.008	0.672 ± 0.030‡§

Each significant QTL detected (LOD > 4.5) is shown with the corresponding linkage group (LG), maximum LOD score and percentage phenotypic variance explained (PVE). *Pitx1* is the most distal marker on linkage group 7, and substantially increased both the LOD and PVE scores compared with the previous most distal marker (*Stn82*). Mean length and height measures were calculated for progeny that inherited two marine (MM), one benthic and one marine (MB), or two benthic alleles (BB) at a marker closely linked to a given QTL (left side measure (mm)/standard body length (mm) ± standard error). Asymmetry is the ratio of left spine length to total combined spine length (left plus right). Statistical analysis was done using one-way analysis of variance. NA, not applicable.

*Significantly different from MM mean ($P < 0.03$).
†Significantly different from MM mean ($P < 0.0001$).
‡Significantly different from MB mean ($P < 0.003$).
§Significantly different from MB mean ($P < 0.0001$).

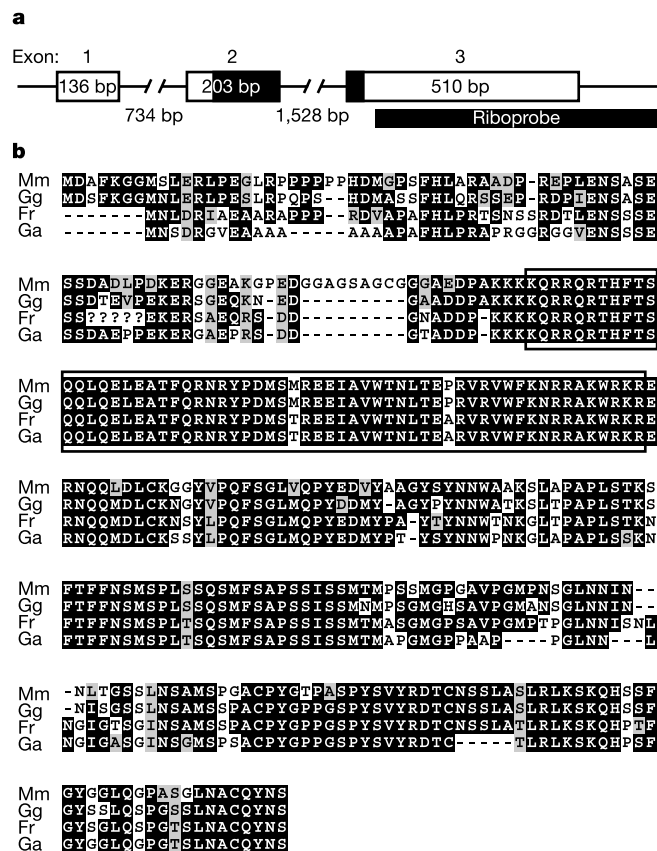


Figure 2 Structure and sequence of the stickleback *Pitx1* locus. **a**, The 849-bp *Pitx1* coding region comprises three exons. The homeodomain is indicated in black, and the sequence used for riboprobe is shown. **b**, Comparison of *Pitx1* amino acid sequences for mouse (*Mus musculus*; Mm), chicken (*Gallus gallus*; Gg), pufferfish (*Fugu rubripes*; Fr) and threespine stickleback (*G. aculeatus*; Ga). Identical amino acids are shaded black whereas conservative substitutions are shaded grey. The homeodomain is boxed. Sequence ambiguities in the predicted pufferfish sequence are denoted by a question mark. Amino acid sequence and splicing of the *Pitx1* transcript is identical in marine and Paxton benthic sticklebacks with complete and reduced pelvises.

opment in the mouse, this gene is highly expressed in hindlimbs but not in forelimbs, giving rise to the alternative descriptive name backfoot²⁸. Null mutations in the gene lead to selective reduction of hindlimbs but not forelimbs. Pelvic limb reduction in *Pitx1* mouse mutants also shows strong directional asymmetry, with greater reduction on the right than on the left side^{25,27}. This effect is probably due to partial functional compensation by the closely related gene *Pitx2* (ref. 27), which is expressed preferentially on the left side during development, a laterality that is highly conserved in mammals, frogs and fish²⁹. To test for similar morphological directional asymmetry in the cross between marine and Paxton benthic sticklebacks, we measured lengths of both left and right pelvic structures, and scored asymmetry as the ratio of the length of the spine on the left to the combined length of both sides. We observed strong directional asymmetry in the F₂ progeny, with longer spines on the left than right side in 78% of animals with asymmetric development (97 out of 124), and a mean ratio of left to total spine length of 0.719 ($P < 0.0001$, one-sided *t*-test). Mapping pelvic asymmetry as a quantitative trait showed strong linkage to the *Pitx1* locus on the distal end of linkage group 7 (LOD = 31.5, Table 1 and Fig. 1a). Directional asymmetry of pelvic reduction in sticklebacks may thus be a naturally occurring phenocopy of the hindlimb-specific directional asymmetry previously seen in laboratory-generated *Pitx1* knockout mice.

Pitx1 sequence comparison in marine and benthic fish

To test whether populations with pelvic reduction have changes in the coding regions of *Pitx1*, we determined the exon/intron structure of the gene (Fig. 2a), and sequenced the entire coding region and exon–intron junctions in both marine and Paxton benthic individuals. The three exons of the stickleback *Pitx1* gene encode a 283-amino-acid protein that shows extensive sequence identity to *Pitx1* sequences previously reported from other fish, birds and mammals (Fig. 2b). No coding region mutations were found that would alter the amino acid sequence of the gene product in the pelvic-reduced Paxton benthic population relative to the marine population. Reverse-transcriptase-mediated polymerase chain reaction (RT–PCR) and sequencing studies confirmed that both marine and Paxton benthic fish correctly assembled the three exons into an intact, spliced *Pitx1* transcript (data not shown).

Altered *Pitx1* gene expression in pelvic-reduced fish

To test for possible regulatory changes in the *Pitx1* locus, we

examined the spatial patterns of expression of *Pitx1* during normal development of both marine and Paxton benthic larvae. In marine fish, *Pitx1* was clearly expressed in bilateral patches of mesenchymal cells just posterolateral to the anterior margin of the ventral median fin, at the site where the pelvic fin bud normally develops (Fig. 3). In addition, *Pitx1* expression was detected at several other sites including the developing thymus, olfactory pits, sensory neuromasts on the head, trunk and tail, and the ventral portion of the developing caudal fin (Fig. 4 and data not shown). Larvae from the Paxton benthic population also showed abundant *Pitx1* expression in the developing thymus, olfactory pits and neuromasts of the head, trunk and tail (Fig. 4 and data not shown). However, no *Pitx1* expression was detectable in the prospective pelvic region (Fig. 3) of Paxton benthic larvae, and expression in the developing caudal fin was also much weaker than in the marine population (Fig. 4). Note that the caudal fin is present in both marine and Paxton benthic larvae, and no significant changes in caudal fin ray number are seen in F_2 progeny that inherit different alleles at the *Pitx1* locus (data not shown). The altered fin expression pattern of *Pitx1* thus cannot be due solely to absence of some structures in Paxton benthic fish. Furthermore, lack of expression in the prospective pelvic region was not due to a developmental delay or difference in timing, as marine larvae continued to show robust expression of *Pitx1* in the pelvic region for several days, whereas no expression was ever detected in the corresponding region of Paxton benthic larvae (Fig. 3 and data not shown). Paxton benthic fish thus show an altered pattern of *Pitx1* expression in some tissues but not others, consistent with a regulatory mutation that disrupts expression in both the prospec-

tive pelvic region and caudal fin.

Parallel evolution of pelvic reduction

Pelvic reduction has evolved multiple times in widely separated lakes throughout the Northern Hemisphere^{8–12}. To test whether the genetic basis of pelvic loss is similar in independent locations, we established a complementation cross between pelvic-reduced fish from Paxton Lake, British Columbia, and Lake Víflsstadavatn, Iceland. The Víflsstadavatn population showed pronounced pelvic reduction, with greater reduction on the right than left side (mean pelvic scores using a standard 0 (all pelvic elements absent) to 4 (all elements present) scoring system¹⁶: 2.78 (left) and 2.12 (right); distribution of scores, 0 = 1.5%, 1 = 12.6%, 2 = 18.6%, 3 = 41.2%, 4 = 26.1% (left) and 0 = 11.1%, 1 = 24.7%, 2 = 16.7%, 3 = 34.8%, 4 = 12.6% (right), $n = 199$). The individual fish chosen for the complementation cross had pelvic scores of 0 (Paxton benthic) and 1 (Víflsstadavatn) on both sides. All 39 F_1 hybrid fish from these parents also showed highly reduced pelvic development: 14 fish had no trace of a pelvic skeleton, 9 had bilateral ovoid rudiments, 15 had a small pelvic rudiment on the left side only and 1 had a small rudiment on the right side only (mean pelvic scores: 0.61 (left), 0.26 (right)). Therefore, the pelvic reduction alleles in these two populations failed to complement and restore normal pelvic morphology. In contrast, a control cross between a Víflsstadavatn pelvic-reduced fish (pelvic score of 1 on each side) and a Little Campbell River marine fish with a robust pelvis (pelvic score of 4 on each side) produced F_1 hybrid fish that all showed strong bilateral development of both pelvic spines and girdles

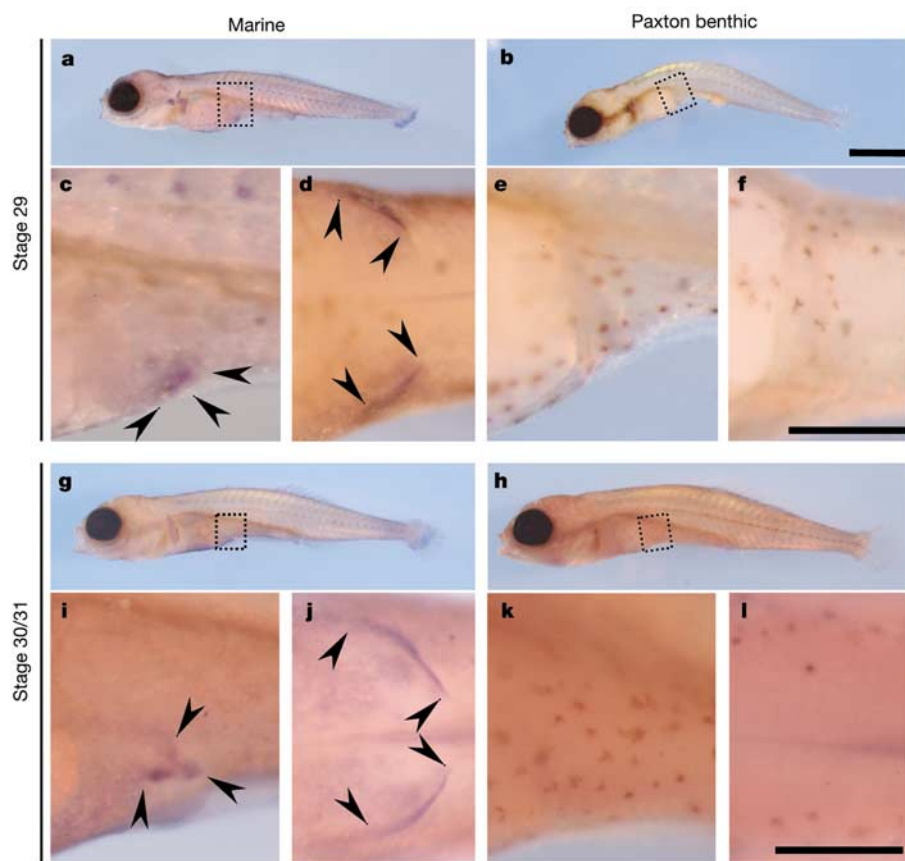


Figure 3 *Pitx1* is expressed in the prospective pelvic region of marine but not Paxton benthic sticklebacks. **a**, Whole-mount *in situ* hybridization shows *Pitx1* expression in the prospective pelvic region of stage 29 marine larvae. Arrowheads mark limits of expression in enlarged lateral (**c**) and ventral (**d**) views. **b**, **e**, **f**, Conversely, no *Pitx1* expression was

detected in the corresponding regions of Paxton benthic larvae. Similar results are seen at later developmental stages (stage 30/31) of marine (**g**, **i**, **j**) and Paxton benthic (**h**, **k**, **l**) larvae. Scale bars: 1 mm (**a**, **b**, **g**, **h**); 0.25 mm (**c**–**f**, **i**–**l**).

(mean pelvic scores: 4.0 (left), 4.0 (right), $n = 41$). These findings indicate that failure of pelvic development in the cross between Paxton and Vífilstadvatn fish was not due to dominant genetic changes in the Vífilstadvatn population. The directional asymmetry in the Icelandic population, and the failure to restore pelvic development in crosses with Paxton benthic fish, suggest that pelvic reduction in these two distinct Pacific and Atlantic basin populations results from defects in similar genes.

Discussion

We have identified a major chromosome region controlling loss of pelvic structures in a natural population of sticklebacks. The absence of recombination between *Pitx1* and the major pelvic-reduction phenotype, the strong directional asymmetry seen in pelvic-reduced fish, and the altered pattern of *Pitx1* expression seen during normal development all suggest that *cis*-acting regulatory mutations in *Pitx1* are a major cause of pelvic reduction in this rapidly evolving system (Fig. 5). Notably, complete inactivation of the *Pitx1* gene in mice leads to severe craniofacial abnormalities, pituitary defects and neonatal lethality^{25,26}. Similarly, null mutations in other known limb-identity genes such as *Tbx4* (ref. 30) and *Tbx5* (refs 31, 32) also cause embryonic lethality due to the pleiotropic roles of these genes in development of other tissues. In contrast, Paxton benthic fish show no alterations in *Pitx1* coding sequences, and changes in gene regulation that disrupt expression only at specific sites in developing larvae. Regulatory mutations in key developmental control genes may provide a general mechanism to selectively alter expression in specific structures and yet preserve expression at other sites required for viability^{33–35}. Similar site-specific regulatory mutations are thought to underlie microevolutionary changes in trichomes, larval hairs, pigmentation and wing eyespot size of fruitflies and butterflies^{36–39}, and muscle mass

variation in pigs⁴⁰. Our study suggests that regulatory mutations in key developmental control genes may also be responsible for major and rapid morphological changes in limb and fin structures during vertebrate evolution.

In this and many other cases where evolutionary changes have been traced to regulatory alterations in major developmental control genes, the actual DNA sequences responsible for tissue-specific expression differences are still unknown^{36–39}. Regulatory mutations are much more difficult to identify at the molecular level than coding region mutations, and could be located at large distances either 5' or 3' to the gene of interest. Comparative analysis of mouse and human genome sequences shows that the *Pitx1* gene is flanked by large regions of potential regulatory sequences, including more than 300 kilobases of flanking DNA that is highly enriched in conserved non-coding sequences⁴¹. These sequences probably include multiple regulatory elements controlling *Pitx1* gene expression in different tissues (Fig. 5c). A major goal of future research will be to identify particular DNA sequences that normally control expression of *Pitx1* in developing hindlimbs (fins), and to compare the structure and function of these regions in marine and Paxton benthic fish.

Are mutations in *Pitx1* likely to underlie other examples of naturally occurring pelvic reduction? Pelvic-reduced populations have evolved repeatedly from marine ancestors in many different locations. The complementation cross between Paxton and Icelandic fish suggests that pelvic reduction has occurred by similar genetic

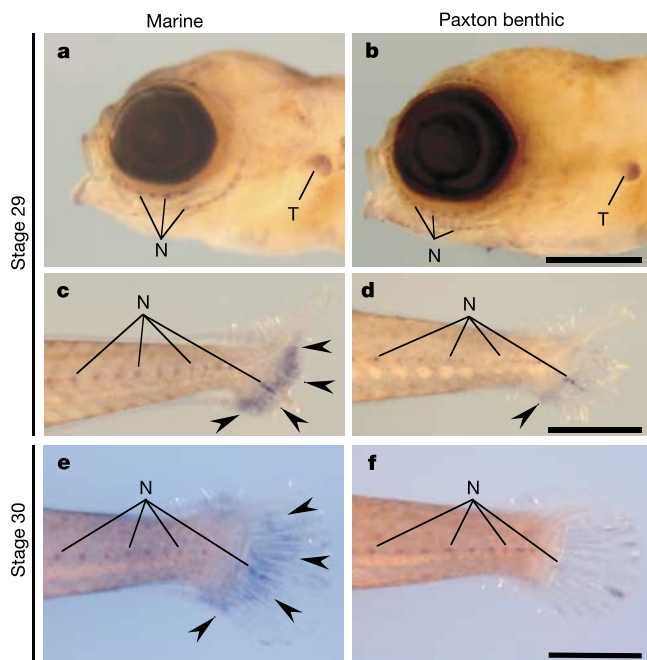


Figure 4 Site-specific regulatory changes of *Pitx1* expression. **a–f**, *Pitx1* is expressed at similar levels in several non-fin tissues of stage 29 and stage 30 marine (**a**, **c**, **e**) and Paxton benthic (**b**, **d**, **f**) larvae, including thymus (T) and cranial and trunk neuromasts (N). At stage 29, however, *Pitx1* expression is considerably higher in the ventral caudal fins (arrowheads) of marine (**c**) than Paxton benthic larvae (**d**). At stage 30, caudal fins undergo normal dorsal rotation in both marine (**e**) and Paxton benthic (**f**) populations, but only the marine population shows expression of *Pitx1* in the caudal fin region. Scale bars: 0.5 mm.

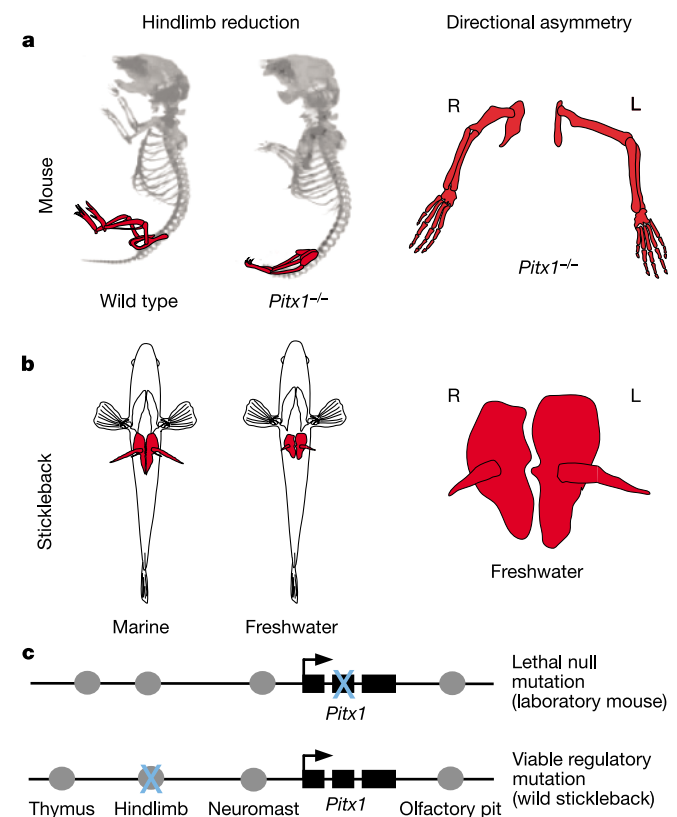


Figure 5 Comparison of pelvic reduction in laboratory and natural populations.

a, Inactivation of *Pitx1* in laboratory mice causes a marked decrease in size of hindlimbs but not forelimbs (modified from refs 25, 27), and greater reduction on the right than left side. **b**, Sticklebacks show similar *Pitx1*-linked selective reduction and asymmetry in hindlimbs. **c**, Null mutations in *Pitx1* in mice disrupt multiple functions and lead to neonatal lethality. In contrast, regulatory mutations in modular *cis*-acting regulatory elements of sticklebacks could produce major morphological changes in particular body regions without disrupting functions in other tissues.

mechanisms in populations located more than 5,700 km apart. Recent experiments also have shown that both *Pitx1* and *Tbx4* fail to be expressed in the prospective pelvic region of pelvic-reduced fish in Scotland⁴². Many other populations show the same left–right asymmetry that is a characteristic feature of *Pitx1*-linked pelvic reduction in mice and Paxton benthic fish, including multiple populations of *Gasterosteus aculeatus*^{13,42,43}, a 10-million-year-old fossil sample of *G. doryssus*¹⁸ and pelvic-reduced populations from the distantly related stickleback genera *Culea*⁴⁴ and *Pungitius*⁴⁵. Mutations in or closely linked to the *Pitx1* locus may contribute to many other examples of evolutionary reduction of pelvic structures in natural populations, a possibility that can now be tested by further genetic studies and direct analysis of *Pitx1* structure and regulation in multiple populations, species and genera. □

Methods

Fish crosses and husbandry

A wild-caught marine female stickleback from Onnechikappu stream (east coast of Hokkaido Island, Japan) was crossed to a wild-caught benthic male from Paxton Lake (British Columbia). Hybrid progeny were raised to maturity in 30-gallon aquaria. F₁ hybrids were mated in pairs, and 2,600 F₂ progeny were raised to a standard length of >30 mm. A total of 375 full siblings from a single F₂ family were used for detailed phenotype and genotype studies. Nineteen fish from this family, and 51 fish from two additional F₂ families, showed complete bilateral absence of pelvic spines and were used for finer recombination mapping of *Pitx1*.

Wild-caught Gjogur marine (Iceland) and Paxton Lake sticklebacks were spawned in the laboratory for *in situ* hybridization studies. Progeny were raised in 15-gallon aquaria at 18 °C.

Complementation crosses were set up between wild-caught, pelvic-reduced fish from Lake Vífilsstaðavatn (Gardabaer, Iceland) and reduced- or full-pelvic fish from Paxton Lake or Little Campbell River (British Columbia). Crosses were raised in 30-gallon aquaria at 18 °C and scored for pelvic phenotypes after 4 months.

Morphological analysis and QTL mapping

Fish were fixed and measured as described²⁰. Pelvic spines were measured from the distal tip to the point of articulation with the pelvis, pelvic girdles from the anterior to posterior tip, and ascending branch from the dorsal tip to the point of articulation with the spine. Asymmetry in the pelvis was analysed as the ratio of left spine length to total combined spine length. Phenotypic measures were analysed with MapQTL 3.0 (ref. 46) as described²⁰, using either raw pelvic measurements, measurements scaled to standard lengths of fish, or residuals of a regression of pelvic measurements on standard body length. Each method led to identification of similar major and minor chromosome regions.

BAC library screening, sequencing and genotyping

A stickleback BAC library (CHORI-213; Children's Hospital Oakland Research Institute) prepared from marine fish (Salmon River, British Columbia) with robust pelvic structures was screened on high-density filters using radiolabelled overgo oligonucleotide probes⁴⁷ designed to conserved *Pitx1*, *Pitx2* and *Tbx4* coding sequences: *Pitx1* screen, 5'-TTTGATCGGCTCCCGTCA-3' and 5'-GTGCCAGTACAACAGCTGACGGGA-3', 5'-CGCTCAGACTCAAGTCCAACAG-3' and 5'-GCCGAAGCTCGGGTGCTGTTTGG-3'; *Pitx1/Pitx2* combined screen, 5'-CTGTCCACCAAGAAGCTTCACTTTC-3' and 5'-GGCTCATGGAGTTGAAGAAAGTGA-3', 5'-CTTCAACTCCATGAGCCCTCTTAC-3' and 5'-GAACATGGACTGAGATGTAAGAGG-3', 5'-CTTCACTTCTTCAACATCCATGAG C-3' and 5'-GACTGAGATGTAAGAGGGCTCATGG-3', 5'-TGGAGTTGAAGAAGTGAA-3' and 5'-CCGCTGTCTCCACCAAGAAGCTTCACTTTC-3'; *Tbx4* screen, 5'-TGGATCCCAAGAACACGGCTACTG-3' and 5'-GTGGAAGACATGGGTACATGAGG-3', 5'-ATCCAAGAACACGGCTACTGTA-3' and 5'-CGTGAAGACATGGGTACAGTAG-3'. Positive BAC clones were secondarily screened using PCR with gene-specific primers: *Pitx1* and *Pitx2*, 5'-AGCCGTACGAAGACATGTACC-3' and 5'-CATGGTCATGGAGGAGATGG-3'; *Tbx4*, 5'-CCCATTTCGAACCACTTCC-3' and 5'-AGTAATGCTCTCCGAGACC-3'. Clones were end-sequenced using T7 and SP6 primers on an ABI377 or ABI3730xl automatic sequencer, followed by complete (*Pitx1*) or partial (*Pitx2*, *Tbx4*) sequencing of coding regions (GenBank accession numbers AY517634–AY517636). Genetic mapping was carried out with PCR primers that flanked a trinucleotide repeat in intron 1 of *Pitx1* (set 1, 5'-TCGACGAGACTCATCTCACG-3' and 5'-AATGTTTTCGCTCGGTTTCC-3'; set 2, 5'-AAACCGGAGAAGGTAAACG-3' and 5'-GACTGCTCCGTTTGATGAGG-3'), a four-nucleotide polymorphism in the T7 end of *Pitx2* BAC clone 59-O19 (5'-GTCTCTGTCTCAGGCTTGGC-3' and 5'-AAGGCTGTGACTGTAAGAAAGGG-3'; GenBank AY517637), and a dinucleotide repeat in the T7 end of *Tbx4* BAC clone 143-J5 (5'-CTGCTCTCTCTTTTGTACCC-3' and 5'-TCTCTCTTACCACATGAAGGG-3'; GenBank AY517638). Each marker was amplified from genomic DNA of F₂ progeny as described²⁰.

Pitx1 sequence comparisons

PCR primers flanking exons 1, 2 and 3 were used to amplify *Pitx1* coding regions from Paxton Lake fish (Exon 1, 5'-GAGGGCAATGTTTACTCAGC-3' and 5'-CATCTTCCC TGACAGACTGC-3'; Exon 2, 5'-GTGCTGGAAGAAGCTTCAAGG-3' and 5'-CAGCGT GACTAATAGAAGGG-3'; Exon 3, 5'-ACCTTTGTTGGTAAATCCGC-3' and

5'-ACACAAAACCCGCATAATCC-3'). Products were cloned into pCR2.1-TOPO (Invitrogen), sequenced, and aligned using Sequencher 4.1 software.

RT-PCR

Complementary DNA was synthesized (RETROScript, Ambion) using total RNA from marine and Paxton benthic fish. A 756-base-pair (bp) fragment of the *Pitx1* transcript that included the 3' portion of exon 1, all of exons 2 and 3, and 6 bp of 3' untranslated region was amplified and sequenced as above (primers 5'-CCGCGTTTCACTTACCAGAG-3' and 5'-CTCCCGTCAGCTGTTGTACTG-3').

In situ hybridization

A 656-bp fragment of *Pitx1* was amplified by PCR from CHORI-213 BAC clone 164-F21 using primers 5'-agaatggatccCCTGTGCAAGAGCAGCTACC-3' and 5'-actattctagaGCGCCTTCATCATAAAAAGC-3', where lower-case letters indicate engineered restriction sites, and upper-case letters indicate *Pitx1*-specific sequence. The amplified fragment was cloned into the pCR4-TOPO cloning vector (Invitrogen), and riboprobe synthesis was performed as described⁴⁸.

For whole-mount *in situ* hybridization, marine and Paxton benthic larvae were staged under a dissecting microscope using morphology of the head, and caudal, dorsal and anal fins⁴⁹, fixed overnight at 4 °C in 4% paraformaldehyde in phosphate buffered saline (PBS; Gibco), and stored at –20 °C in methanol. Rates of larval development were slightly variable within and among clutches. Stage 29 marine larvae were fixed at 14 days post-fertilization (d.p.f.), Paxton benthics at 18 d.p.f.; stage 30/31 marine larvae at 20 d.p.f., Paxton benthics at 29 d.p.f. *In situ* hybridizations were performed essentially as described⁴⁸. After signal development in BCIP/NBT (SigmaFast tablets, Sigma), the alkaline phosphatase reaction was halted by washing in 100 mM glycine in PBT (PBS plus 0.1% Tween-20), followed by dehydration with two 5-min washes in 100% ethanol, and re-hydration into PBT. Larvae were post-fixed in 4% paraformaldehyde and photographed in PBT.

Received 13 November 2003; accepted 30 January 2004; doi:10.1038/nature02415.

1. Darwin, C. *The Origin of Species* 450–456 (John Murray, London, 1859).
2. Struthers, J. On the bones, articulations, and muscles of the rudimentary hind-limb of the Greenland right-whale (*Balaena mysticetus*). *J. Anat. Phys.* **15**, 141–176, 302–321 (1881).
3. Bejder, L. & Hall, B. K. Limbs in whales and limbliness in other vertebrates: mechanisms of evolutionary and developmental transformation and loss. *Evol. Dev.* **4**, 445–458 (2002).
4. Lande, R. Evolutionary mechanisms of limb loss in tetrapods. *Evolution* **32**, 73–92 (1978).
5. Cohn, M. J. & Tickle, C. Developmental basis of limbliness and axial patterning in snakes. *Nature* **399**, 474–479 (1999).
6. Hoogland, R. D., Morris, D. & Tinbergen, N. The spines of sticklebacks (*Gasterosteus* and *Pygosteus*) as means of defense against predators (*Perca* and *Esox*). *Behaviour* **10**, 205–230 (1957).
7. Reimchen, T. E. Structural relationships between spines and lateral plates in threespine stickleback (*Gasterosteus aculeatus*). *Evolution* **37**, 931–946 (1983).
8. Bell, M. A. Reduction and loss of the pelvic girdle in *Gasterosteus* (Pisces): a case of parallel evolution. *Nat. Hist. Mus. LA Contrib. Sci.* **257**, 1–36 (1974).
9. Moodie, G. E. E. & Reimchen, T. Phenetic variation and habitat differences in *Gasterosteus* populations of the Queen Charlotte Islands. *Syst. Zool.* **25**, 49–61 (1976).
10. Campbell, R. N. & Williamson, R. B. The fishes of inland waters in the Outer Hebrides. *Proc. R. Soc. Edinb.* **77B**, 377–393 (1979).
11. Edge, T. A. & Coad, B. W. Reduction of the pelvic skeleton in the three-spined stickleback *Gasterosteus aculeatus* in 2 lakes of Quebec Canada. *Can. Field-Nat.* **97**, 334–336 (1983).
12. Bell, M. A. Interacting evolutionary constraints in pelvic reduction of threespine sticklebacks, *Gasterosteus aculeatus* (Pisces, Gasterosteidae). *Biol. J. Linn. Soc.* **31**, 347–382 (1987).
13. Reimchen, T. E. Spine deficiency and polymorphism in a population of *Gasterosteus aculeatus*—an adaptation to predators. *Can. J. Zool.* **58**, 1232–1244 (1980).
14. Reist, J. D. Predation upon pelvic phenotypes of brook stickleback, *Culaea inconstans*, by selected invertebrates. *Can. J. Zool.* **58**, 1253–1258 (1980).
15. Giles, N. The possible role of environmental calcium levels during the evolution of phenotypic diversity in Outer-Hebridean populations of the three-spined stickleback, *Gasterosteus aculeatus*. *J. Zool.* **199**, 535–544 (1983).
16. Bell, M. A., Orti, G., Walker, J. A. & Koenings, J. P. Evolution of pelvic reduction in threespine stickleback fish—a test of competing hypotheses. *Evolution* **47**, 906–914 (1993).
17. Zhiuganov, V. V. & Zotin, A. A. Pelvic girdle polymorphism and reproductive barriers in the ninespine stickleback *Pungitius pungitius* (L.) from northwest Russia. *Behaviour* **132**, 1095–1105 (1995).
18. Bell, M. A., Baumgartner, J. V. & Olson, E. C. Patterns of temporal change in single morphological characters of a Miocene stickleback fish. *Paleobiology* **11**, 258–271 (1985).
19. McPhail, J. D. Ecology and evolution of sympatric sticklebacks (*Gasterosteus*): evidence for a species pair in Paxton Lake, Texada Island, British Columbia. *Can. J. Zool.* **70**, 361–369 (1992).
20. Peichel, C. L. *et al.* The genetic architecture of divergence between threespine stickleback species. *Nature* **414**, 901–905 (2001).
21. Crow, J. F. & Kimura, M. *An Introduction to Population Genetics Theory* 182–183 (Burgess, Minneapolis, 1970).
22. Gibson-Brown, J. J. *et al.* Evidence of a role for T-box genes in the evolution of limb morphogenesis and the specification of forelimb/hindlimb identity. *Mech. Dev.* **56**, 93–101 (1996).
23. Logan, M., Simon, H. G. & Tabin, C. Differential regulation of T-box and homeobox transcription factors suggests roles in controlling chick limb-type identity. *Development* **125**, 2825–2835 (1998).
24. Logan, M. & Tabin, C. J. Role of *Pitx1* upstream of *Tbx4* in specification of hindlimb identity. *Science* **283**, 1736–1739 (1999).
25. Lancôt, C., Moreau, A., Chamberland, M., Tremblay, M. L. & Drouin, J. Hindlimb patterning and mandible development require the *Ptx1* gene. *Development* **126**, 1805–1810 (1999).
26. Szeto, D. P. *et al.* Role of the Bicoid-related homeodomain factor *Pitx1* in specifying hindlimb morphogenesis and pituitary development. *Genes Dev.* **13**, 484–494 (1999).
27. Marciel, A., Dumontier, E., Chamberland, M., Camper, S. A. & Drouin, J. *Pitx1* and *Pitx2* are required for development of hindlimb buds. *Development* **130**, 45–55 (2003).

28. Shang, J., Luo, Y. & Clayton, D. A. Backfoot is a novel homeobox gene expressed in the mesenchyme of developing hind limb. *Dev. Dyn.* **209**, 242–253 (1997).
29. Campione, M. *et al.* The homeobox gene Pitx2: mediator of asymmetric left-right signaling in vertebrate heart and gut looping. *Development* **126**, 1225–1234 (1999).
30. Naiche, L. A. & Papaioannou, V. E. Loss of Tbx4 blocks hindlimb development and affects vascularization and fusion of the allantois. *Development* **130**, 2681–2693 (2003).
31. Bruneau, B. G. *et al.* A murine model of Holt-Oram syndrome defines roles of the T-box transcription factor Tbx5 in cardiogenesis and disease. *Cell* **106**, 709–721 (2001).
32. Garrity, D. M., Childs, S. & Fishman, M. C. The heartstrings mutation in zebrafish causes heart/fin Tbx5 deficiency syndrome. *Development* **129**, 4635–4645 (2002).
33. Carroll, S. B. Endless forms: the evolution of gene regulation and morphological diversity. *Cell* **101**, 577–580 (2000).
34. Stern, D. L. Evolutionary developmental biology and the problem of variation. *Evolution* **54**, 1079–1091 (2000).
35. Tautz, D. Evolution of transcriptional regulation. *Curr. Opin. Genet. Dev.* **10**, 575–579 (2000).
36. Stern, D. L. A role of Ultrabithorax in morphological differences between *Drosophila* species. *Nature* **396**, 463–466 (1998).
37. Sucena, E. & Stern, D. L. Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby. *Proc. Natl Acad. Sci. USA* **97**, 4530–4534 (2000).
38. Kopp, A., Duncan, I., Godt, D. & Carroll, S. B. Genetic control and evolution of sexually dimorphic characters in *Drosophila*. *Nature* **408**, 553–559 (2000).
39. Beldade, P., Brakefield, P. M. & Long, A. D. Contribution of Distal-less to quantitative variation in butterfly eyespots. *Nature* **415**, 315–318 (2002).
40. van Laere, A. S. *et al.* A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* **425**, 832–836 (2003).
41. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
42. Cole, N. J., Tanaka, M., Prescott, A. & Tickle, C. Expression of limb initiation genes and clues to the morphological diversification of threespine stickleback. *Curr. Biol.* **13**, R951–R952 (2003).
43. Bell, M. A., Francis, R. C. & Havens, A. C. Pelvic reduction and its directional asymmetry in threespine sticklebacks from the Cook Inlet region, Alaska. *Copeia* **1985**, 437–444 (1985).
44. Nelson, J. S. Evidence of a genetic basis for absence of pelvic skeleton in brook stickleback, *Culaea inconstans*, and notes on geographical distribution and origin of loss. *J. Fish. Res. Board Can.* **34**, 1314–1320 (1977).
45. Nelson, J. S. Absence of pelvic complex in ninespine sticklebacks, *Pungitius pungitius*, collected in Ireland and Wood Buffalo National Park region, Canada, with notes on meristic variation. *Copeia* **1971**, 707–717 (1971).
46. van Ooijen, J. W. & Maliepaard, C. *MapQTL Version 3.0: Software for the Calculation of QTL Positions and Genetic Maps* (Centre for Plant Breeding and Reproductive Research, Wageningen, 1996).
47. Ross, M. T., LaBrie, S., McPherson, J. & Stanton, V. P. in *Current Protocols in Human Genetics* (ed. Dracopoli, N.) (Wiley, New York, 1999).
48. Wilkinson, D. G. in *Essential Developmental Biology: A Practical Approach* (eds Stern, C. D. & Holland, P. W. H.) 257–274 (IRL Press, Oxford, 1993).
49. Swarup, H. Stages in the development of the stickleback *Gasterosteus aculeatus* (L.). *J. Embryol. Exp. Morphol.* **6**, 373–383 (1958).
50. Van Ooijen, J. W. LOD significance thresholds for QTL analysis in experimental populations of diploid species. *Heredity* **83**, 613–624 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank S. Mori for providing marine sticklebacks; E. Elfarðóttir, G. I. Gudbrandsson, L. E. Jónsdóttir, S. Ingólfssdóttir and K. Einarsson for collecting Icelandic sticklebacks and scoring the parental population from Víflsstadavatn; and F. Chan and members of the Kingsley laboratory for discussions. This work was supported in part by grants from the National Institutes of Health (to D.M.K.), the Natural Sciences and Engineering Research Council of Canada and the Canada Foundation for Innovation (to D.S.), and the City of Gardabaer (to B.J.); a Helen Hay Whitney Foundation postdoctoral fellowship (to M.D.S.); and a Howard Hughes Medical Institute predoctoral fellowship (to M.E.M.). D.S. is a Canada Research Chair, and D.M.K. is an Associate Investigator of the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.M.K. (kingsley@cmgm.stanford.edu). The sequences described in this paper have been deposited in GenBank under accession numbers AY517634–AY517638.

Formation of massive black holes through runaway collisions in dense young star clusters

Simon F. Portegies Zwart¹, Holger Baumgardt², Piet Hut³, Junichiro Makino⁴ & Stephen L. W. McMillan⁵

¹Astronomical Institute 'Anton Pannekoek', and Institute for Computer Science, University of Amsterdam, Kruislaan 403, The Netherlands

²Institute of Advanced Physical and Chemical Research RIKEN, 2-1 Hirosawa, Wako-shi, Saitama 351-019, Japan

³Institute for Advanced Study, Princeton, New Jersey 08540, USA

⁴Department of Astronomy, University of Tokyo, Tokyo 113, Japan

⁵Department of Physics Drexel University, Philadelphia, Pennsylvania 19104, USA

A luminous X-ray source is associated with MGG 11—a cluster of young stars ~200 pc from the centre of the starburst galaxy M 82 (refs 1, 2). The properties of this source are best explained^{3,4} by invoking a black hole with a mass of at least 350 solar masses (350 $M_{\odot}$), which is intermediate between stellar-mass and supermassive black holes. A nearby but somewhat more massive cluster (MGG 9) shows no evidence of such an intermediate-mass black hole^{1,3}, raising the issue of just what physical characteristics of the clusters can account for this difference. Here we report numerical simulations of the evolution and motion of stars within the clusters, where stars are allowed to merge with each other. We find that for MGG 11 dynamical friction leads to the massive stars sinking rapidly to the centre of the cluster, where they participate in a runaway collision. This produces a star of 800–3,000 $M_{\odot}$, which ultimately collapses to a black hole of intermediate mass. No such runaway occurs in the cluster MGG 9, because the larger cluster radius leads to a mass segregation timescale a factor of five longer than for MGG 11.

Using the Keck NIRSPEC spectrometer, McCrady *et al.*⁵ have made accurate measurements of the bulk parameters of MGG 11 and MGG 9, two of the brightest star clusters in the central region of M 82. Figure 1 shows X-ray³ and near-infrared⁵ images of the area of interest around the two clusters. The relative positional accuracies of both sets of observations are better than 1 arcsec. However, the absolute pointing accuracy is much poorer, for both telescopes. Although apparently off-centre, the positions of the star cluster MGG 11 and the bright X-ray source are in fact consistent with one another (D. Pooley, personal communication).

MGG 9 and MGG 11 have quite similar ages, in the range 7–12 Myr (ref. 5). The line-of-sight velocity dispersion of MGG 11 ($\sigma_r = 11.4 \pm 0.8 \text{ km s}^{-1}$) is somewhat smaller than that of MGG 9 ($\sigma_r = 15.9 \pm 0.8 \text{ km s}^{-1}$). Combining these numbers with the projected half-light radii, $r = 1.2 \text{ pc}$ for MGG 11 and $r = 2.6 \text{ pc}$ for MGG 9, McCrady *et al.* estimate a total cluster mass of about $(3.5 \pm 0.7) \times 10^5 M_{\odot}$ for MGG 11, and about four times higher for MGG 9.

In young dense clusters like MGG 11, supermassive stars may form through repeated collisions^{6–8}. The collision rate will be greatly enhanced if massive stars have time to reach the core before exploding as supernovae⁹. Dynamical friction implies a characteristic timescale t_{df} for a massive star in a roughly circular orbit to sink from the half-mass radius R to the cluster centre¹⁰:

$$t_{\text{df}} \approx \frac{\langle m \rangle}{100 M_{\odot}} \frac{0.138 N}{\ln(0.11 M / 100 M_{\odot})} \left(\frac{R^3}{GM} \right)^{1/2} \quad (1)$$

Here $\langle m \rangle$ and M are the mean stellar mass and the total mass of the cluster, respectively, N is the number of stars, and G is the gravitational constant. For definiteness, we have evaluated t_{df} for a 100 $M_{\odot}$ star. Less-massive stars undergo weaker dynamical friction, and thus

must start at smaller radii in order to reach the cluster centre on a similar timescale.

For MGG 11, we find $t_{\text{df}} \approx 3 \text{ Myr}$, which is comparable to the main-sequence lifetimes of the most massive stars. On the other hand, for MGG 9, $t_{\text{df}} \approx 15 \text{ Myr}$. Thus, massive stars in MGG 11 can easily reach the centre of the cluster before exploding as supernovae, whereas those in MGG 9 do not. Given the high central density of MGG 11, its massive stars, once accumulated in the cluster centre, cannot avoid a runaway collision⁶.

We have tested this scenario by carrying out star-by-star simulations of MGG 11 and MGG 9 using two independently developed N -body codes, Starlab^{11,12} and NBODY4 (refs 13, 14). A more extensive discussion of those simulations is presented as Supplementary Information. Briefly, we choose the initial conditions of our model clusters so that today, at an age of 7–12 Myr, they have mass functions, luminosities, half-mass radii and velocity dispersions in agreement with the McCrady *et al.* observations. As we do not know either the initial or the current central densities of either cluster, the concentration c is treated as a free parameter controlling the initial central density of our model clusters. (The concentration parameter is defined as the logarithm of the ratio of the core radius to the tidal radius: $c \equiv \log(R_t/R_c)$.)

We find that, for $c > 2$ (which for 'King'¹⁵ models is equivalent to a dimensionless central potential $W_0 \geq 9$), our MGG 11 models do indeed show runaway growth via repeated collisions. Figure 2 presents, for several such simulations, the growth in mass of the star that would ultimately become the most massive star in the cluster. We will refer to this star simply as 'the runaway star'. On the basis of detailed supernova calculations, we assume that stars having masses $> 260 M_{\odot}$ collapse to black holes without significant mass loss in supernova explosions¹⁶. Our stellar evolution models for

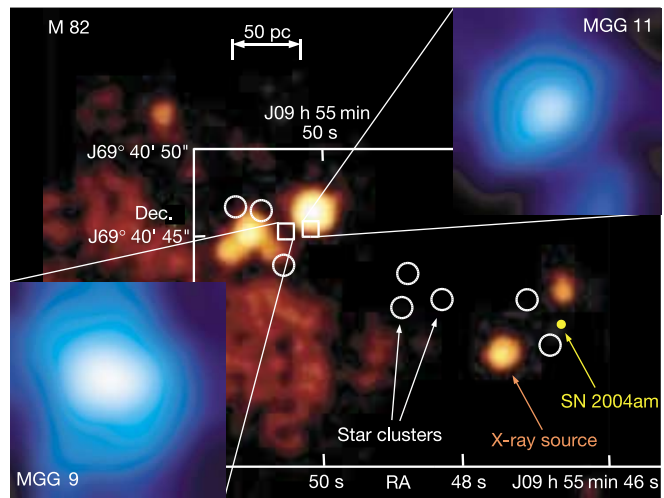


Figure 1 Chandra X-ray image of the relevant part of the starburst galaxy M 82 with the observed star clusters indicated. The main colour image (red–yellow–white) is from the 28 October 1999 X-ray observation by Matsumoto *et al.*³, with about arcsecond positional accuracy. The brightest X-ray source (M 82 X-1) is near the centre of the image. The star clusters, from table 3 of McCrady *et al.*⁵, are indicated by circles. The positions of the two star clusters MGG 9 and MGG 11 are indicated with squares. The magnified infrared images (dark blue–pale blue–white) of these star clusters from the McCrady *et al.* observations are presented in the upper right (MGG 11) and lower left (MGG 9) insets. The type IIp supernova 2004am discovered on 6 March is indicated with a yellow dot to the right of the image²⁷. A recently discovered 54.4 ± 0.9 mHz quasi-periodic oscillator is not shown because of its low (7 arcsec) positional accuracy⁴, but its position is consistent with the X-ray source in MGG 11. A millimetre source²⁸ roughly centred around the two clusters is not shown either, because it is a large shell-like structure with a diameter of ~14'' × 9''. RA, right ascension; dec., declination.

stars with masses between 50 and $1,000 M_{\odot}$ are based on detailed calculations for such high-mass stars^{17,18}.

The quantitative differences evident in Fig. 2 between the simulations performed with Starlab and those using NBODY4 are due mainly to the different radii assumed for the runaway star in those two packages. Stars of masses $\geq 100 M_{\odot}$ in NBODY4 are larger (in a time-averaged sense) by about a factor of 3 compared with those in Starlab. (More details are provided in Supplementary Information.) Because gravitational focusing dominates the collision cross-section, this difference propagates linearly in the collision rate, explaining the factor ~ 3 difference in the final mass of the runaway star. Apart from this effect, we find that our qualitative results are quite insensitive to the details of the adopted evolution prescription.

For MGG 11, the runaway star typically experienced a total of ~ 10 – 100 collisions. Most of them occurred during the first 3 Myr; that is, before the star became a black hole. The collision counterparts are usually 30 – $50 M_{\odot}$ main-sequence stars. By the time the runaway star collapsed to a black hole it had reached a mass of 800 – $3,000 M_{\odot}$. Later the black hole may capture a companion star to become an ultraluminous X-ray source¹⁹.

No episode of runaway growth was seen in our MGG 11 models with $c < 2$, nor in any of the MGG 9 simulations, regardless of initial concentration. Thus we see very clearly that differences in bulk parameters, specifically the dynamical friction timescale for the most massive stars, can readily explain why MGG 11 might host an intermediate-mass black hole (IMBH) whereas MGG 9 does not. Simulations of MGG 11 with $c > 2$ ($W_0 \geq 9$) reached core collapse²⁰ before 1 Myr. Models with $c < 2$ ($W_0 \leq 8$) also showed some increase in central density, but the maximum density was much lower (and occurred later) because the collapse was stopped by

supernova mass loss after ~ 3 Myr. Thus, a collision runaway could not occur in these clusters.

Figure 3 summarizes the results of our simulations, illustrating how both high initial concentration and short dynamical friction timescales are needed to lead to a collision runaway. In addition to the parameters for MGG 9 and MGG 11, we also indicate on the figure the location of the young star cluster [W99]1 (ref. 21), one of several star clusters in the Antennae system for which accurate structure parameters have been determined²². Its parameters are similar to those of MGG 9. The cluster [W99]1 has a bright X-ray point source (luminosity $L_{0.2-10 \text{ keV}} \approx 10^{38.7} \text{ erg s}^{-1}$) as a counterpart²³, the luminosity of which is consistent with a 'normal' high-mass X-ray binary containing a magnetized neutron star or a stellar-mass black hole²³. Star clusters such as MGG 11, MGG 9 and [W99]1 are richly populated with black holes and neutron stars—on the basis of our adopted mass function, we expect these clusters to contain some 1,400 stellar-mass black holes and up to about 1,000 neutron stars. The formation of an ordinary high-mass X-ray binary is therefore not surprising.

The relevant parameters of five other young clusters in the Antennae²² all fall far to the right of Fig. 3 and are therefore not expected to contain IMBHs. This is consistent with the absence of ultraluminous X-ray sources in these clusters. The Milky Way contains (at least) four young dense star clusters, of which the Arches, Quintuplet and NGC 3603 have the right conditions for multiple stellar collisions to occur. However, the relatively small number of stars in these systems may prevent the growth of an object massive enough to collapse to an IMBH. In any case, these clusters are currently too young to have experienced any supernovae. The slightly older star cluster Westerlund 1 (ref. 24) is sufficiently massive, and fulfils the criteria for producing an

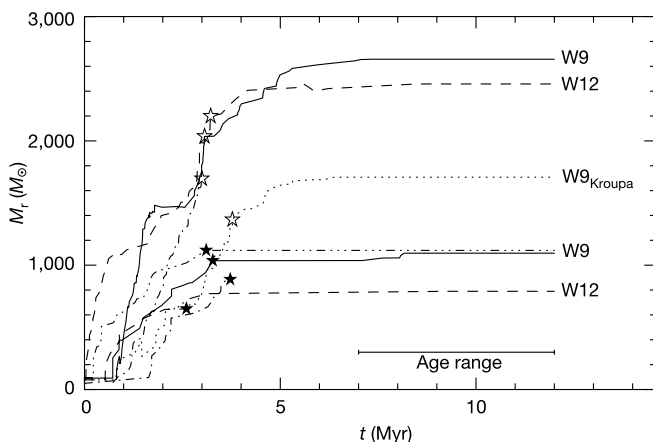


Figure 2 The growth in mass of the collision runaway star with time. The choice of initial concentration is labelled by the central potential, where W12(9) implies $W_0 = 12(9)$. For both choices, the top curves give the NBODY4 results, and the bottom curves the Starlab results. The runaway masses in NBODY4 are larger, because that code adopts larger stellar radii, as discussed in Supplementary Information. The star symbols indicate the moment when the runaway experiences a supernova, typically around 3 Myr. The open and filled stars indicate simulations performed with NBODY4 and Starlab, respectively. The solid and dashed curves show M_t for a Salpeter²⁹ initial mass function (IMF) with a lower limit of $1 M_{\odot}$ and $c \approx 2.1$ ($W_0 = 9$) and $c \approx 2.7$ ($W_0 = 12$). The dash-dotted curves are for two models with $W_0 = 9$ with an upper limit to the IMF of $50 M_{\odot}$, instead of the standard $100 M_{\odot}$ used in the other calculations; we terminated these runs at the moment the runaway star experiences a supernova. The dash-three-dotted curve shows the result for $W_0 = 12$ with a Salpeter IMF and with 10% primordial binaries. Finally, the dotted curve shows results for $W_0 = 9$ and a Kroupa³⁰ IMF with a minimum mass of $0.1 M_{\odot}$, in a stimulation with 585,000 stars. The observed age range of MGG 11 and MGG 9 is indicated by the horizontal bar.

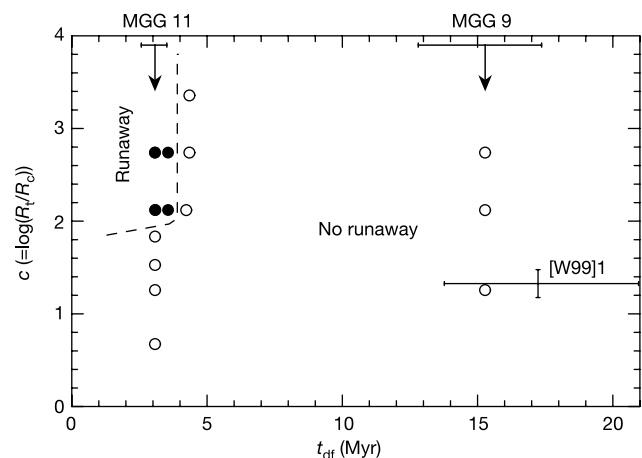


Figure 3 The area of parameter space for which runaway collision can occur, and where the process is prevented. Conditions for runaway merging are identified through our simulations in the $\{t_{\text{dif}}, c\}$ plane, where t_{dif} is the dynamical friction timescale for $100 M_{\odot}$ stars, and c the cluster concentration parameter. The horizontal error bars for the initial conditions for MGG 9 and MGG 11 reflect errors in the observed cluster mass and projected half-light radius⁵. The best fit for the concentration parameter of the Antennae star cluster [W99]1 is indicated by a vertical bar near $t_{\text{dif}} \approx 17$ Myr; the horizontal error bar reflects the uncertainty in the measured cluster mass and radius²². Filled circles indicate simulations resulting in runaway merging, while open circles correspond to simulations in which no runaway merging occurred. The evolution of the mass of the runaway for the two leftmost filled circles are presented in Fig. 2, for a variety of initial conditions. Star clusters in the upper left corner enclosed by the dashed line are expected to host an IMBH formed by the runaway growth of a single star. In the other parts of the diagram (lower concentration and/or larger dynamical friction time), an IMBH cannot form by this process. R_t , tidal radius; R_c , core radius.

IMBH via the process described above. So far, no bright X-ray source has been found in this cluster.

On the basis of theoretical considerations, supported by extensive numerical simulations, we have shown that an IMBH of mass $\sim 800\text{--}3,000M_{\odot}$ is expected to form through runaway collisions in the star cluster MGG 11, but not in MGG 9. The requirement for the formation of such an IMBH in MGG 11 is that the cluster was born with $c \geq 2$ and $t_{\text{df}} \leq 4\text{ Myr}$. Although high by the standards of typical open and globular clusters observed in our Galaxy, such a density is not uncommon among young ($\lesssim 3\text{ Myr}$ old) star clusters. Examples include NGC 3603 ($c \geq 2.08$)²⁵ in the Milky Way, and R 136 in the 30 Doradus region of the Large Magellanic Cloud²⁶. We propose that the IMBH thus formed is the origin of the ultraluminous $L_{0.2\text{--}10\text{ keV}} \approx 10^{41}\text{ erg s}^{-1}$ X-ray point source seen at the position of MGG 11. $\square$

Received 22 October 2003; accepted 1 March 2004; doi:10.1038/nature02448.

- Matsumoto, H. & Tsuru, T. G. X-ray evidence of an AGN in M82. *Publ. Astron. Soc. Jpn* **51**, 321–331 (1999); Erratum. *Publ. Astron. Soc. Jpn* **51**, 567 (1999).
- Kaaret, P. *et al.* Chandra high-resolution camera observations of the luminous X-ray source in the starburst galaxy M82. *Mon. Not. R. Astron. Soc.* **321**, L29–L32 (2001).
- Matsumoto, H. *et al.* Discovery of a luminous, variable, off-center source in the nucleus of M82 with the Chandra high-resolution camera. *Astrophys. J.* **547**, L25–L28 (2001).
- Strohmayer, T. E. & Mushotsky, R. F. Discovery of X-ray quasi-periodic oscillations from an ultraluminous X-ray source in M82: Evidence against beaming. *Astrophys. J.* **586**, L61–L64 (2003).
- McCrady, N., Gilbert, A. M. & Graham, J. R. Kinematic masses of super star clusters in M82 from high-resolution near-infrared spectroscopy. *Astrophys. J.* **596**, 240–252 (2003).
- Portegies Zwart, S. F., Makino, J., McMillan, S. L. W. & Hut, P. Star cluster ecology. III. Runaway collisions in young compact star clusters. *Astron. Astrophys.* **348**, 117–126 (1999).
- Portegies Zwart, S. F. & McMillan, S. L. W. The runaway growth of intermediate-mass black holes in dense star clusters. *Astrophys. J.* **576**, 899–907 (2002).
- Gürkan, M. A., Freitag, M. & Rasio, F. A. Formation of massive black holes in dense star clusters. I. Mass segregation and core collapse. *Astrophys. J.* (in the press); preprint at (<http://arxiv.org/list/astro-ph/0308449>) (2003).
- Ebisuzaki, T. *et al.* Missing link found? The “runaway” path to supermassive black holes. *Astrophys. J.* **562**, L19–L22 (2001).
- Spitzer, L. Jr & Hart, M. H. Random gravitational encounters and the evolution of spherical systems. I. Method. *Astrophys. J.* **164**, 399–409 (1971).
- Portegies Zwart, S. F., McMillan, S. L. W., Hut, P. & Makino, J. Star cluster ecology-IV. Dissection of an open star cluster: photometry. *Mon. Not. R. Astron. Soc.* **321**, 199–226 (2001).
- Hut, P., Makino, J., McMillan, S. L. W. & Portegies Zwart, S. *Starlab* (<http://manybody.org/manybody/starlab.html>) (1995).
- Aarseth, S. J. From NBODY1 to NBODY6: The growth of an industry. *Publ. Astron. Soc. Pacif.* **111**, 1333–1346 (1999).
- Baumgardt, H. & Makino, J. Dynamical evolution of star clusters in tidal fields. *Mon. Not. R. Astron. Soc.* **340**, 227–246 (2003).
- King, I. R. The structure of star clusters. III. Some simple dynamical models. *Astron. J.* **71**, 64–75 (1966).
- Heger, A., Fryer, C. L., Woosley, S. E., Langer, N. & Hartmann, D. H. How massive single stars end their life. *Astrophys. J.* **591**, 288–300 (2003).
- Stothers, B. & Chin, C.-W. On two theories of the cyclical outbursts of eta Carinae. *Astrophys. J.* **489**, 319–330 (1997).
- Ishii, M., Munetaka, U. & Kato, M. Core-halo structure of a chemically homogeneous massive star and bending of the zero-age main sequence. *Publ. Astron. Soc. Jpn* **51**, 417–424 (1999).
- Hopman, C., Portegies Zwart, S. F. & Alexander, T. Ultraluminous X-ray sources as intermediate mass black holes fed by tidally captured stars. *Astrophys. J. Lett.* (submitted); preprint at (<http://arxiv.org/list/astro-ph/0312597>) (2003).
- Heggie, D. C. & Hut, P. *The Gravitational Million-Body Problem* (Cambridge Univ. Press, Cambridge, 2003).
- Whitmore, B. C. *et al.* The luminosity function of young star clusters in “The Antennae” galaxies. *Astrophys. J.* **118**, 1551–1576 (1999).
- Mengel, S., Lehnert, M. D., Thattai, N. & Genzel, R. Dynamical masses of young star clusters in NGC 4038/4039. *Astron. Astrophys.* **383**, 137–152 (2002).
- Zezas, A., Fabbiano, G., Rots, A. H. & Murray, S. S. Chandra observations of “The Antennae” galaxies (NGC 4038/4039). III. X-ray properties and multiwavelength associations of the X-ray source population. *Astrophys. J.* **577**, 710–725 (2002).
- Piatti, A. E., Bica, E., Clariá, J. J. Fundamental parameters for the highly reddened young open clusters Westerlund 1 and 2. *Astron. Astrophys. Suppl.* **127**, 423–432 (1998).
- Moffat, A. F. J., Drissen, L. & Shara, M. M. NGC 3603 and its Wolf-Rayet stars: Galactic clone of R136 at the core of 30 Doradus, but without the massive surrounding cluster halo. *Astrophys. J.* **436**, 183–193 (1994).
- Campbell, B. *et al.* Hubble Space Telescope planetary camera images of R136. *Astron. J.* **104**, 1721–1742 (1992).
- Singer, D., Pugh, H. & Li, W. Supernovae 2004ak, 2004al, and 2004am. *IAU Circ. No.* 8297 (2004).
- Matsumoto, M. S. *et al.* Formation of a massive black hole at the center of the superbubble in M82. *Astrophys. J.* **545**, L107–L111 (2000).
- Salpeter, E. E. The luminosity function and stellar evolution. *Astrophys. J.* **121**, 161–167 (1955).
- Kroupa, P. On the variation of the initial mass function. *Mon. Not. R. Astron. Soc.* **322**, 231–246 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank N. McCrady and D. Pooley for discussions on MGG 11, T. G. Tsuru for an accurate position of M82 X-1, and E. van den Heuvel for critically reading the manuscript. This work was supported by NASA ATP, the Royal Netherlands Academy of Sciences (KNAW), the Dutch Organization of Science (NWO), and the Netherlands Research School for Astronomy (NOVA).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.F.P.Z. (spz@science.uva.nl).

Chaotic electron diffusion through stochastic webs enhances current flow in superlattices

T. M. Fromhold, A. Patané, S. Bujkiewicz, P. B. Wilkinson, D. Fowler, D. Sherwood, S. P. Stapleton, A. A. Krokhin*, L. Eaves, M. Henini, N. S. Sankeshwar* & F. W. Sheard

School of Physics and Astronomy, University of Nottingham, Nottingham NG7 2RD, UK

* Present addresses: Center for Nonlinear Science, University of North Texas, PO Box 311427, Denton, Texas 76203, USA, and Instituto de Física, Universidad Autónoma de Puebla, Puebla 72570, Mexico (A.A.K.); Department of Physics, Karnatak University, Dharwad 580 003, India (N.S.S.)

Understanding how complex systems respond to change is of fundamental importance in the natural sciences. There is particular interest in systems whose classical newtonian motion becomes chaotic^{1–22} as an applied perturbation grows. The transition to chaos usually occurs by the gradual destruction of stable orbits in parameter space, in accordance with the Kolmogorov–Arnold–Moser (KAM) theorem^{1–3,6–9}—a cornerstone of nonlinear dynamics that explains, for example, gaps in the asteroid belt². By contrast, ‘non-KAM’ chaos switches on and off abruptly at critical values of the perturbation frequency^{6–9}. This type of dynamics has wide-ranging implications in the theory of plasma physics¹⁰, tokamak fusion¹¹, turbulence^{6,7,12}, ion traps¹³, and quasicrystals^{6,8}. Here we realize non-KAM chaos experimentally by exploiting the quantum properties of electrons in the periodic potential of a semiconductor superlattice^{22–27} with an applied voltage and magnetic field. The onset of chaos at discrete voltages is observed as a large increase in the current flow due to the creation of unbound electron orbits, which propagate through intricate web patterns^{6–10,12–16} in phase space. Non-KAM chaos therefore provides a mechanism for controlling the electrical conductivity of a condensed matter device: its extreme sensitivity could find applications in quantum electronics and photonics.

Superlattices (SLs) comprise alternating layers of different semiconductor materials, which form a nanometre-scale periodic potential and energy bands, known as minibands, for electron motion perpendicular to the layers^{22–27} (Fig. 1a, b). Electrons can be accelerated through the energy-crystal momentum dispersion curve, $E(p_x)$, of the first miniband (lower curve in Fig. 1c) by using a bias voltage V to create an electric field F along the SL axis. At low V , the electrons accelerate slowly and scatter when they are still near the miniband bottom. In this regime, the electron drift (average) velocity v_d increases linearly with increasing F (see black curve in Fig. 1d). But at high V , electrons can reach the top of the miniband before scattering and therefore perform semiclassical Bloch oscillations^{23,24,26} (owing to repeated Bragg reflection of the electron waves) of angular frequency $\omega_B = eFd/\hbar$, where d is the SL period and $-e$ is the electron charge. The onset of Bloch oscillations

localizes the electrons perpendicular to the layers, thereby causing ν_d to decrease with increasing F (refs 23, 24; black curve in Fig. 1d). By using a tilted magnetic field $\mathbf{B}$ to couple Bloch oscillations to cyclotron orbits in the plane of the layers, we have created and studied the effects of non-KAM chaos in a real semiconductor device. At discrete voltages, the coupling transforms the localized stable Bloch trajectories into unbounded chaotic electron paths. This striking signature of non-KAM chaos produces strong resonant enhancement of ν_d (arrowed peaks in the green curve of Fig. 1d), and also has interesting quantum-mechanical manifestations, in particular an abrupt transition from fully localized to highly extended energy eigenfunctions.

Figure 1a shows the SL used in our experiments. An unusual feature is the InAs monolayer at the centre of each GaAs/AlAs quantum well, which lowers the energy of the first miniband,

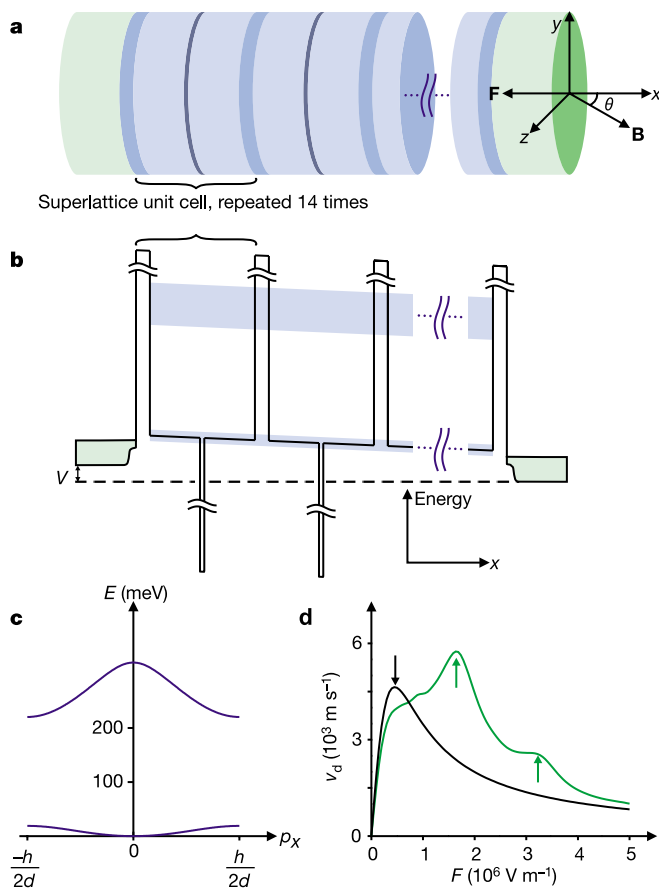


Figure 1 Properties of the SL. **a**, Schematic diagram of the SL, whose unit cell comprises two 3.5-nm-thick GaAs layers (light blue), a 0.3-nm (one monolayer)-thick InAs layer (dark blue), and a 1-nm-thick AlAs barrier layer (mid blue)²⁵. Each layer has a Si doping concentration $n_D = 3 \times 10^{16} \text{ cm}^{-3}$. The structure contains 14 unit cells, enclosed by 50-nm-thick GaAs ohmic contacts (green) with $n_D = 10^{17} \text{ cm}^{-3}$, and 0.6- μm -thick GaAs outer layers (not shown) heavily doped with $n_D = 2 \times 10^{18} \text{ cm}^{-3}$. The layers are grown by molecular beam epitaxy and processed into circular 25- μm -diameter mesas. Coordinate axes show orientation of $\mathbf{F}$ and $\mathbf{B}$. **b**, Schematic variation of the electronic potential energy with position x normal to the layers, for $V \neq 0$. The quantum wells produce a periodic potential (only two complete unit cells are shown for clarity) supporting two minibands (blue)²³. Green areas represent electron gases in the contacts. **c**, Energy versus crystal momentum dispersion curves for the two minibands. In contrast to previous systems used to study quantum chaos^{2–5,13,14,16–21}, the effective classical hamiltonian, equation (1) originates from the intrinsically quantum-mechanical nature of the dispersion curves. **d**, Plots of ν_d versus F calculated for $B = 11 \text{ T}$ with $\theta = 0$ (black curve: arrow marks peak) and $\theta = 45^\circ$ (green curve: arrows mark additional peaks).

thereby facilitating electron injection from the emitter contact and reducing inter-miniband Zener tunnelling²⁵. Figure 2a shows current–voltage $I(V)$ characteristics measured when $\mathbf{B}$, of magnitude 11 T, is applied at a range of tilt angles θ relative to the SL (x) axis (Fig. 1a). When $\theta = 0$ (lowest curve in Fig. 2a), I increases monotonically with increasing V below $V_P \approx 200 \text{ mV}$, but thereafter has an almost constant value $I_P \approx 11 \text{ mA}$. The bend in $I(V)$, marked by the arrow in Fig. 2a, originates from the onset of Bloch electron localization^{23,24}. For $10^\circ \leq \theta \leq 55^\circ$, the slope of each $I(V)$ curve increases when V reaches $\sim 250 \text{ mV}$. Above this threshold voltage, there is a region of enhanced current flow, which produces strong resonant peaks in the differential conductance $G = dI/dV$ versus V plots (Fig. 2b). The peak amplitude decreases as θ approaches 0, where the small remnant peak (arrowed in Fig. 2b) is a Stark-cyclotron resonance²⁷, originating from weak elastic scattering between the localized quantum states produced by $\mathbf{B}$ and $\mathbf{F}$.

The stronger resonant peaks observed in tilted $\mathbf{B}$ have a fundamentally different origin: non-KAM classical chaos of miniband electrons. To investigate the electron dynamics, we solved Hamilton's equations for the semiclassical hamiltonian¹⁵

$$H(x, z, p_x, p_z) = E(p_x) + \frac{(q_y^2 + p_z^2)}{2m^*} - eFx \quad (1)$$

where $E(p_x)$ is the lower non-parabolic dispersion curve in Fig. 1c, and the electronic effective mass in the y - z plane, m^* , is 0.067 times

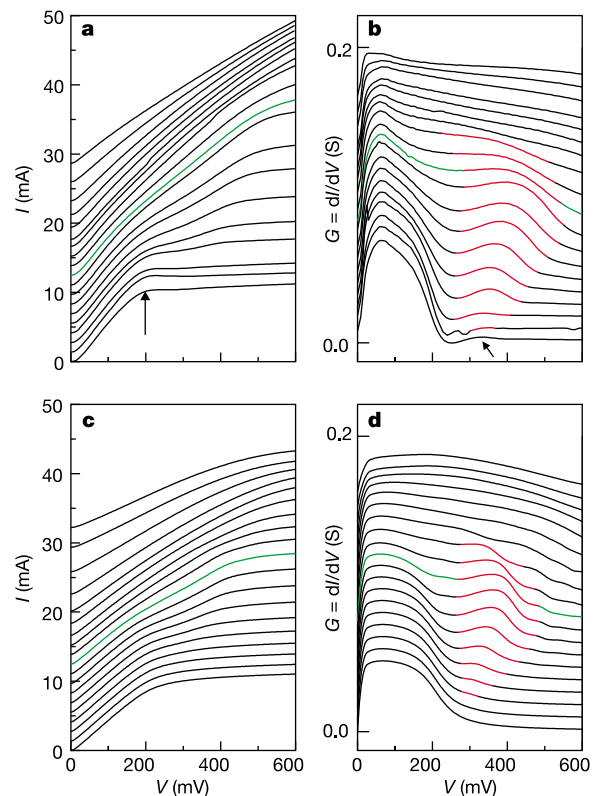


Figure 2 Resonant enhancement of current. **a**, Experimental $I(V)$ curves measured for $B = 11 \text{ T}$ and $\theta = 0$ (bottom curve) to 90° (top curve) at 5° intervals at a lattice temperature of 4.2 K. For $10^\circ \leq \theta \leq 55^\circ$, each curve contains a region of enhanced I beyond $V \approx 250 \text{ mV}$. **b**, Differential conductance plots of the data in **a** reveal strong resonant peaks (red). **c**, Theoretical $I(V)$ characteristics (for same parameters as **a**). **d**, Differential conductance plots of the traces in **c**. Curves in **a–d** are offset vertically for clarity and those for $\theta = 45^\circ$ are green. In theory and experiment, the resonant peaks in $G(V)$ initially shift slightly to higher V as θ increases, because the enhanced conductance leads to a higher electron charge density in the SL, which increases F and V (refs 22, 26).

the free electron mass. The in-plane kinetic momentum components are p_z and $q_y = eB(x \sin \theta - z \cos \theta) = (\omega_C \cos \theta)^{-1} dp_z/dt$, where the cyclotron frequency $\omega_C = eB/m^*$ (see ref. 15). It can be shown that the momentum component p_z satisfies the equation of motion

$$\frac{d^2 p_z}{dt^2} + (\omega_C \cos \theta)^2 p_z = - \sum_{n=1}^{\infty} \frac{na_n m^* \omega_C^2 d \sin 2\theta}{2\hbar} \sin[n(Kp_z - \omega_B t - \phi)] \quad (2)$$

corresponding to a one-dimensional simple harmonic oscillator

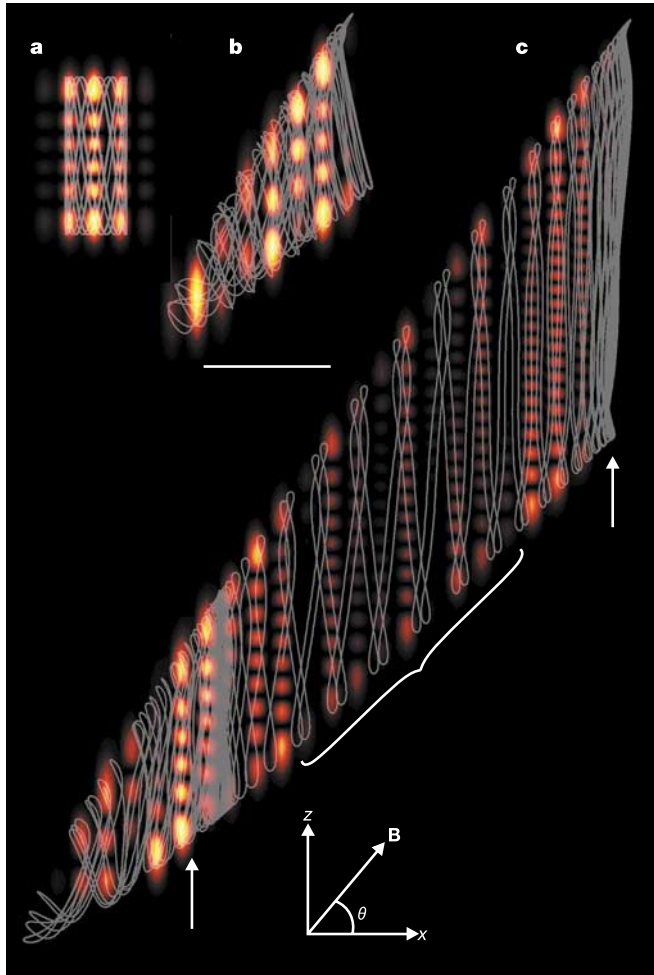


Figure 3 Electron trajectories and wavefunctions. Grey translucent curves: classical electron trajectories in the x - z plane (axes inset) overlaid on corresponding plots of $|\Psi(x, z)|^2$ (black zero, yellow high) at $B = 11$ T. **a**, When $\theta = 0$, the probability distribution is concentrated within the turning points of the classical orbit. For $\theta \approx 15^\circ$, **b** strongly couples electron motion along x and z , thereby producing chaotic trajectories and irregular wavefunctions, as in **b** and **c**. **b**, Off resonance, here for $r = (1 + \sqrt{5})/4$ and $\theta = 50^\circ$, the trajectories and wavefunctions are localized. **c**, On resonance, here for $r = 1$ and $\theta = 50^\circ$, the wavefunctions extend across many SL periods in correspondence with the unbounded trajectories. Wavefunction **c** is shaped by the stochastic web shown in Fig. 4f. A region of high probability density (yellow peaks), associated with a concentration of orbital loops, occurs when the electron is trapped on the first (inner) ring-shaped web filament (lower left-hand arrow marks x value corresponding to this ring), and is therefore unable to progress through the SL. But when the electron transfers onto the quasi-linear filaments, it shifts rapidly along x , following widely-spaced orbital loops (within bracket), which correspond to low probability density. The wavefunction is bounded on the right by the second ring-shaped web filament (right-hand arrow marks x position corresponding to this ring), which impedes electron flow. Scale bar, 5d.

driven by time (t)-dependent plane waves¹⁵. In equation (2), $K = -d \tan \theta / \hbar$, $\phi = d[p_x(t=0) + p_z(t=0) \tan \theta] / \hbar$ and the miniband Fourier coefficients a_n are defined by

$$E(p_x) = a_0 - \sum_{n=1}^{\infty} a_n \cos \frac{np_x d}{\hbar} \quad (3)$$

where a_0 is chosen so that $E(0) = 0$. In our SL, the coupling between adjacent quantum wells is sufficiently weak that only the $n = 1$ terms ($a_1 = 9.5$ meV) need to be considered in equations (2) and (3)^{15,25}. Remarkably, for $0 < \theta < 90^\circ$, the electron motion depends only on equation (2), because the solution p_z uniquely determines all of the other dynamical variables. Consequently, the semiclassical dynamics of a miniband electron in tilted **B** reduce to a harmonic oscillator driven by a monochromatic plane wave: a non-KAM chaotic system that has been the focus of intense theoretical study, but never realized in an actual quantum system that can be explored experimentally^{6–16}.

We now consider electron trajectories obtained by numerical solution of equation (2). When $\theta = 0$, the plane wave has zero amplitude and the motion along the x - and z -directions is separable. Electrons undergo Bloch oscillations along x and cyclotron motion about **B** (Fig. 3a). Tilting **B** produces nonlinear coupling of the Bloch and cyclotron motion, which causes electrons injected into the miniband to follow chaotic trajectories. These trajectories usually remain localized as in Fig. 3b. But, at discrete voltages for which $r = \omega_B / (\omega_C \cos \theta)$ is an integer, the chaotic orbits change discontinuously from localized to unbounded (Fig. 3c). This resonant delocalization can be understood by considering the Poincaré sections (slices through phase space^{1–3}) shown in Fig. 4a–c and enlarged in Fig. 4d–f. Of crucial importance for understanding these sections is that as an electron's x -coordinate increases, it produces points further from the centre of the section¹⁵. When $\theta = 0$ (Fig. 4a, d), each electron orbit is confined to a single ring in phase space and therefore to a particular region of the x -axis, where it performs Bloch oscillations. Tilting **B** generates a chaotic sea (Fig. 4b, e), which is enclosed by isolated stable islands resembling the ring pattern at $\theta = 0$. Orbits within the chaotic sea are extended (Fig. 3b), but remain bounded by the encircling stable islands. However, when $0 < \theta < 90^\circ$, the phase space structure changes dramatically at discrete F values for which r is an integer. When this resonance condition is satisfied, the electron orbits map out intricate 'stochastic web' patterns^{6–10,12–16} (Fig. 4c, f). Quasi-linear web filaments link the rings that are isolated from one another off resonance. This delocalizes the electrons and therefore increases I . Moving F infinitesimally off resonance destroys the stochastic web by removing the quasi-linear filaments (as in Fig. 4b, e), thereby localizing the electrons and reducing I . Consequently, stochastic web formation gives rise to an unusual type of metal–insulator-like transition at a series of discrete voltages where the electron trajectories change from localized (insulator) to unbounded (conductor).

To quantify the effect of electron delocalization, we calculate v_d using the classical kinetic formula^{15,23,26}

$$v_d(F) = \frac{\delta}{\tau} \sum \int_0^\infty v_x(t) \exp(-t/\tau) dt \quad (4)$$

where the summation is over 2,500 starting velocities corresponding to a miniband electron temperature of 100 K, and $v_x(t) = \partial E / \partial p_x$ is the velocity along the SL axis for a given starting condition. The parameters $\delta = [\tau_e / (\tau_e + \tau_i)]^{1/2}$ and $\tau = \delta \tau_i$ depend on the elastic and inelastic scattering times, $\tau_e = 21$ fs and $\tau_i = 1.5$ ps, determined from the measured values of V_p and I_p (ref. 26). When $\theta = 0$, the $v_d(F)$ curve (black trace in Fig. 1d) exhibits the Esaki-Tsu peak²³ at $F = 5 \times 10^5$ V m⁻¹ (marked by black arrow). Tilting **B** produces additional v_d peaks at F values for which $r = 1$ and 2 (marked by left- and right-hand green arrows in Fig. 1d), when the electrons undergo rapid diffusive motion through stochastic webs.

We used the $v_d(F)$ curves to make self-consistent drift-diffusion²⁶ calculations of the $I(V)$ characteristics (Fig. 2c). For $15^\circ \leq \theta \leq 55^\circ$, the slope of each calculated $I(V)$ curve increases sharply at $V \approx 250$ mV. This generates strong resonant peaks in the theoretical $G(V)$ characteristics (Fig. 2d), in good agreement with those observed in experiment (Fig. 2b). Each peak occurs at the voltage for which $r = 1$ throughout most of the SL layers, so that the phase space is threaded by an infinite stochastic web. Electrons undergo rapid diffusive motion through the web, and hence through the SL itself, thereby generating the resonant peak in $G(V)$. We have observed $r = 1$ (and also $r = 2$) resonances in three different SLs for $6\text{ T} \leq B \leq 20\text{ T}$. We emphasize that chaos-assisted transport through stochastic webs is a particularly strong effect that is clearly visible in the undifferentiated $I(V)$ curves (see, for example, the trace for $\theta = 25^\circ$ in Fig. 2a, which exhibits a 50% increase in the current flow over the voltage range corresponding to the resonant peak in $G(V)$). The current resonances shown in Fig. 2 for a lattice temperature of 4.2 K are also clearly visible at room temperature, although they are slightly broadened. This indicates that in our SL devices, raising the temperature has little effect on the inherent sensitivity of non-KAM chaos. Experimental observation of the

current resonances confirms that stochastic web formation imposes large additional peaks in the $v_d(F)$ curves of electrons in the SL, at F values that can be controlled by changing B and/or θ . The ability to tailor the $v_d(F)$ curves in this way could be useful for the development of tunable high-frequency (GHz to THz) oscillators²⁶.

The quantized energy eigenfunctions of the SL, $\Psi(x, z)$, also change discontinuously from fully localized (Fig. 3a, b) to highly extended when the resonance condition is satisfied (Fig. 3c). To relate the eigenfunctions directly to the classical phase space, we use $\Psi(x, z)$ to calculate Wigner functions, $W(p_z, q_y)$, which are quantum-mechanical analogues of the Poincaré sections in Fig. 4a–f^{23,19}. When $\theta = 0$ (Fig. 4g), $W(p_z, q_y)$ has maximal amplitude around the ring in classical phase space (green in Fig. 4d) produced by the orbit whose cyclotron energy equals the quantized energy for motion along z . When $\theta = 50^\circ$, the off resonance Wigner function (Fig. 4h) is concentrated within the chaotic sea at the section centre (Fig. 4e). But on resonance (Fig. 4i), the Wigner function reveals filaments of the stochastic web in Fig. 4f (it lies entirely within the second ring-shaped filament), plus stable islands enclosed by those filaments¹⁴. So, remarkably, the quantum eigenfunctions of this three-dimensional time-independent system

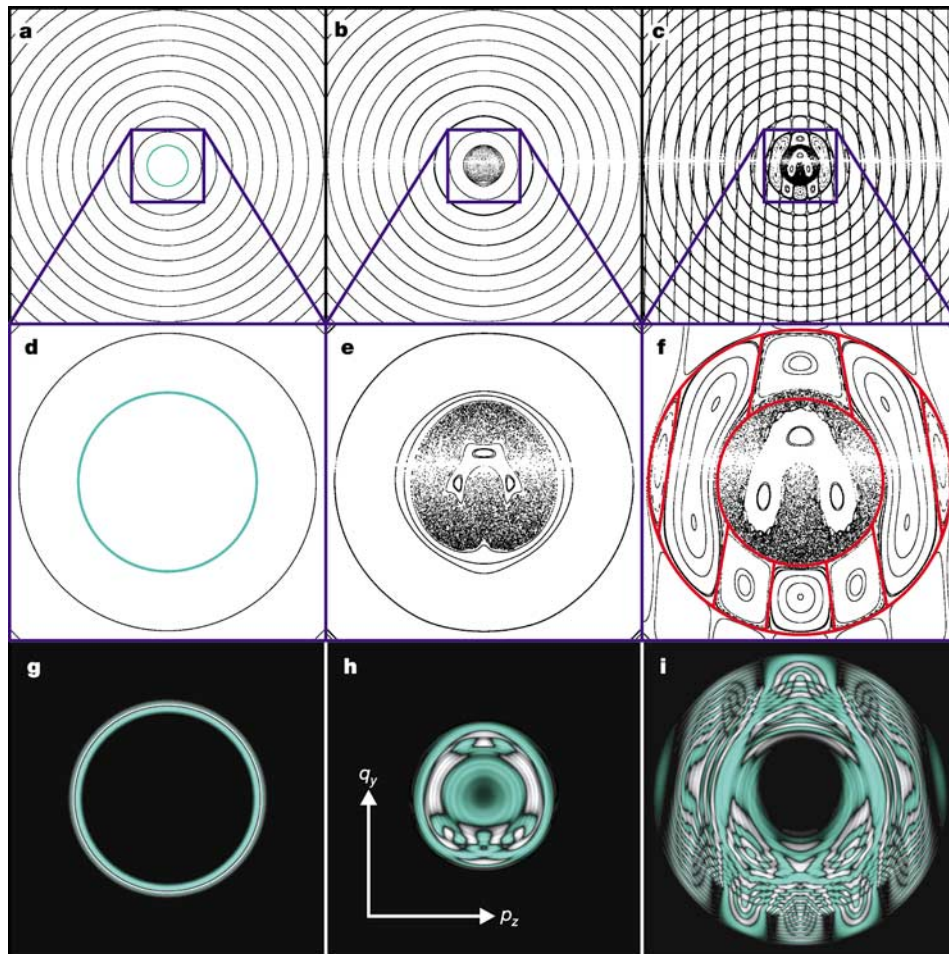


Figure 4 Classical and quantum phase space structure. **a–f**, Poincaré sections calculated from electron orbits with total energy $H = 0$, see equation (1), at $B = 11\text{ T}$. Points show $q_y \propto dp_z/dt$ (vertical) and p_z (horizontal) whenever $p_x = 0$. From equation (1), this condition ensures that the distance of each point from the centre of each section $(0, 0)$ is $\propto \sqrt{x}$, where x is the electron's position along the SL axis. **a**, When $\theta = 0^\circ$, the points lie on rings corresponding to cyclotron motion about **B**. **b**, When $\theta \geq 15^\circ$, a chaotic sea forms at the centre of the section, here for $r = (1 + \sqrt{5})/4$ (off resonance) and $\theta = 50^\circ$. **c**, On resonance, here for $r = 1$ at $\theta = 50^\circ$, filaments extending from the chaotic sea form a stochastic web. Electrons are driven out along the quasi-linear web filaments by

resonant absorption of energy from the plane wave, and so progress rapidly through the SL (Fig. 3c). The electron orbits are unbounded along x because the web extends to infinity^{8,15}. As the electron moves through the SL, potential energy gained from **F** is transferred into kinetic energy for motion in the y - z plane by the tilted **B**-field. **d–f**, Enlargements of regions of **a–c** within blue boxes. **g–i**, Wigner function values (green > 0 , black $= 0$, white < 0) in (p_z, q_y) plane (axes inset in **h**) corresponding to sections **d–f**, and wavefunctions in Fig. 3a–c respectively. Red lines in **f** highlight web filaments. Panels **a–c** (**d–i**) have side lengths of 7×10^{-25} (1.6×10^{-25}) kg m s^{-1} .

are shaped by the phase space structure of a classical one-dimensional periodically driven harmonic oscillator.

The stochastic web pattern depends on the shape of $E(p_x)$, whose Fourier coefficients determine the right-hand side of equation (2). By changing the SL composition, we can tailor $E(p_x)$ to realize stochastic webs with different two-dimensional tiling patterns. Consequently, our work links the electrical properties of nanostructures with the mathematical properties of tilings, including the Penrose tiling²⁸ that can occur in the phase space of a driven harmonic oscillator⁸.

We have realized non-KAM chaos for electrons in a SL with stationary applied electric and magnetic fields that, surprisingly, act like a THz plane wave. At certain critical voltages the electron phase space is threaded by infinite stochastic webs, which delocalize the electron orbits and wavefunctions. This unique feature of non-KAM chaos reveals itself as a strong resonant enhancement of the measured current flow through the SL. Our results suggest that stochastic webs could provide an ultra-sensitive mechanism for controlling transmission through other periodic structures, including ultra-cold atoms in optical lattices^{20,21,29}, and photonic crystals in which a modulation of the lattice constant could bend light rays in a way analogous to the effect of F and B on the electron trajectories in the SL³⁰. Non-KAM chaos, which began with theoretical work, now has an unexpected experimental realization in semiconductor physics. It provides a new method of switching with the potential for developing electronic and photonic devices that exploit the intrinsic sensitivity of chaos. □

Received 21 November 2003; accepted 27 February 2004; doi:10.1038/nature02445.

- Lichtenberg, A. J. & Leiberman, M. A. *Regular and Chaotic Dynamics* (Springer, New York, 1992).
- Gutzwiller, M. C. *Chaos in Classical and Quantum Mechanics* (Springer, New York, 1990).
- Stöckmann, H.-J. *Quantum Chaos: an Introduction* (Cambridge Univ. Press, Cambridge, 1999).
- Berry, M. V. Quantum chaosology. *Proc. R. Soc. Lond. A* **413**, 183–198 (1987).
- Heller, E. J. Bound-state eigenfunctions of classically chaotic Hamiltonian systems—scars of periodic orbits. *Phys. Rev. Lett.* **53**, 1515–1518 (1984).
- Chernikov, A. A., Sagdeev, R. Z., Usikov, D. A., Yu Zakharov, M. & Zaslavsky, G. M. Minimal chaos and stochastic webs. *Nature* **326**, 559–563 (1987).
- Shlesinger, M. F., Zaslavsky, G. M. & Klafter, J. Strange kinetics. *Nature* **363**, 31–37 (1993).
- Zaslavsky, G. M., Sagdeev, R. Z., Usikov, D. A. & Chernikov, A. A. *Weak Chaos and Quasi-Regular Patterns* (Cambridge Univ. Press, Cambridge, 1991).
- Kamenev, D. I. & Berman, G. P. *Quantum Chaos: a Harmonic Oscillator in Monochromatic Wave* (Rinton, Princeton, New Jersey, 2000).
- Chia, P.-K., Schmitz, L. & Conn, R. W. Stochastic ion behaviour in subharmonic and superharmonic electrostatic waves. *Phys. Plasmas* **3**, 1545–1568 (1996).
- Karney, C. F. F. & Bers, A. Stochastic ion heating by a perpendicularly propagating electrostatic wave. *Phys. Rev. Lett.* **39**, 550–554 (1977).
- Beloshapkin, V. V. *et al.* Chaotic streamlines in pre-turbulent states. *Nature* **337**, 133–137 (1989).
- Gardiner, S. A., Cirac, J. I. & Zoller, P. Quantum chaos in an ion trap: the delta-kicked harmonic oscillator. *Phys. Rev. Lett.* **79**, 4790–4793 (1997).
- Demikhovskii, V. Ya., Kamenev, D. I. & Luna-Acosta, G. A. Quantum weak chaos in a degenerate system. *Phys. Rev. E* **59**, 294–302 (1999).
- Fromhold, T. M. *et al.* Effects of stochastic webs on chaotic electron transport in semiconductor superlattices. *Phys. Rev. Lett.* **87**, 046803 (2001).
- Demikhovskii, V. Ya., Izrailev, F. M. & Malyshev, A. I. Manifestation of Arnol'd diffusion in quantum systems. *Phys. Rev. Lett.* **88**, 154101 (2002).
- Jensen, R. V. Quantum chaos. *Nature* **355**, 311–318 (1992).
- Wilkinson, P. B. *et al.* Observation of 'scarred' wavefunctions in a quantum well with chaotic electron dynamics. *Nature* **380**, 608–610 (1996).
- Fromhold, T. M. *et al.* Tunneling spectroscopy of mixed stable-chaotic electron dynamics in a quantum well. *Phys. Rev. B* **65**, 155312 (2002).
- Hensinger, W. K. *et al.* Dynamical tunnelling of ultra-cold atoms. *Nature* **412**, 52–55 (2001).
- Steck, D. A., Oskay, W. H. & Raizen, M. G. Observation of chaos-assisted tunneling between islands of stability. *Science* **293**, 274–278 (2001).
- Amann, A., Schlesner, J., Wacker, A. & Schöll, E. Chaotic front dynamics in semiconductor superlattices. *Phys. Rev. B* **65**, 193313 (2002).
- Esaki, L. & Tsu, R. Superlattice and negative differential conductivity in semiconductors. *IBM J. Res. Dev.* **14**, 61–65 (1970).
- Tsu, R. & Döhler, G. Hopping conduction in a "superlattice". *Phys. Rev. B* **12**, 680–686 (1975).
- Patanè, A. *et al.* Tailoring the electronic properties of GaAs/AlAs superlattices by InAs layer insertions. *Appl. Phys. Lett.* **81**, 661–663 (2002).
- Wacker, A. Semiconductor superlattices: a model system for nonlinear transport. *Phys. Rep.* **357**, 1–111 (2002).
- Canali, L., Lazzarino, M., Sorba, L. & Beltram, F. Stark-cyclotron resonance in a semiconductor superlattice. *Phys. Rev. Lett.* **76**, 3618–3621 (1996).
- Penrose, R. The role of aesthetics in pure and applied mathematical research. *Bull. Inst. Math. Appl.* **10**, 266–271 (1974).
- Scott, R. G., Bujkiewicz, S., Fromhold, T. M., Wilkinson, P. B. & Sheard, F. W. Effects of chaotic

energy-band transport on the quantized states of ultracold sodium atoms in an optical lattice with a tilted harmonic trap. *Phys. Rev. A* **66**, 023407 (2002).

30. Wilkinson, P. B. & Fromhold, T. M. Chaotic ray dynamics in slowly varying two-dimensional photonic crystals. *Opt. Lett.* **28**, 1034–1036 (2003).

Acknowledgements This work was supported by the EPSRC, The Royal Society, CONACyT, and INSA. S.B. is funded by an EU Marie Curie Fellowship and is grateful for support from the Institute of Physics, Wrocław University of Technology, Poland. We are grateful to R. Airey for processing our samples.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.M.F. (mark.fromhold@nottingham.ac.uk).

Observation of rare-earth segregation in silicon nitride ceramics at subnanometre dimensions

Naoya Shibata¹, Stephen J. Pennycook¹, Tim R. Gosnell², Gayle S. Painter³, William A. Shelton⁴ & Paul F. Becher³

¹Condensed Matter Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

²Pixon LLC, 100 North Country Road, Setauket, New York 11733, USA

³Metals and Ceramics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

⁴Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

Silicon nitride (Si_3N_4) ceramics are used in numerous applications because of their superior mechanical properties^{1,2}. Their intrinsically brittle nature is a critical issue, but can be overcome by introducing whisker-like microstructural features^{3,4}. However, the formation of such anisotropic grains is very sensitive to the type of cations used as the sintering additives^{1,2,5}. Understanding the origin of dopant effects, central to the design of high-performance Si_3N_4 ceramics, has been sought for many years. Here we show direct images of dopant atoms (La) within the nanometre-scale intergranular amorphous films typically found at grain boundaries, using aberration corrected Z-contrast scanning transmission electron microscopy. It is clearly shown that the La atoms preferentially segregate to the amorphous/crystal interfaces. First-principles calculations confirm the strong preference of La for the crystalline surfaces, which is essential for forming elongated grains and a toughened microstructure. Whereas principles of micrometre-scale structural design are currently used to improve the mechanical properties of ceramics, this work represents a step towards the atomic-level structural engineering required for the next generation of ceramics.

Brittleness is a limiting factor in the use of ceramics, so approaches to ensure high fracture resistance are actively being sought. In Si_3N_4 based ceramics⁶, this can be attained through 'self-reinforced' microstructures consisting of elongated β - Si_3N_4 grains embedded in a fine-grained matrix. Analogous to whisker-reinforced ceramics, these reinforcing microstructures promote toughening mechanisms (for example, crack bridging, pull out and crack deflection), especially when combined with the appropriate interfacial strengths^{7,8}. This microscopic composite concept is well in hand; however, the next generation of ceramics will need to

be designed at the atomic level to obtain the required improvements in mechanical properties and reliability. To do this, we must understand the role of interfacial interactions, particularly those involving dopant elements. Until now, our understanding of dopant/alloying effects has largely been based on empirical approaches. The present study integrates theory, atomic level characterization and experiment to resolve an important technical issue in ceramics, and take ceramic design to the atomistic scale.

The reinforcement concept is possible in Si_3N_4 because the grains typically exist as elongated hexagonal prisms; this is a result of the preferred crystal habit combined with anisotropic growth, in which the c -axis growth rate exceeds that normal to the prism faces⁹. This anisotropic grain growth behaviour is known to be markedly altered by the addition of some metal oxides, most notably the rare-earth oxides⁵. Thus, selective dopant additions offer the potential for tailoring the mechanical properties through microstructure optimization. Generally, Si_3N_4 ceramics are fabricated through liquid phase sintering, where oxynitride amorphous intergranular films (IGF) with nanometre-scale thickness are formed at grain boundaries¹⁰. Grain growth along the c -axis (length) direction is controlled by diffusion and is quite fast; however, diametrical

growth on the very smooth prism surfaces of the Si_3N_4 grains is typically reaction rate limited. Dopants will influence the growth process at these prism surfaces, but the mechanism has not been well understood. A limited number of studies have been conducted^{11,12}; however, the atomic level details about how the dopants are distributed at the grain/matrix interface versus within the IGF have been extremely difficult to assess, owing to the very small thickness (that is, <2 nm) of the IGF and its amorphous nature.

Recent progress in aberration correction for scanning transmission electron microscopy (STEM) now offers the possibility of characterizing the detailed atomic structures in localized volumes such as the IGFs using a fine probe with a diameter of less than 1 Å (refs 13,14). In the annular dark-field (ADF)-STEM imaging method, the fine electron probe is scanned across a thin specimen, and electrons scattered by each atomic column are collected via an annular detector to form atomic-resolution images. The spatial resolution is critically dependent on the probe size, and can be greatly improved by the aberration correction¹³. Moreover, by using the ADF detector to collect scattered electrons at higher angles, the image intensities correspond directly to the atomic number (Z) of constituent atoms in the materials^{15,16}. In order to investigate the dopant state within the IGF in Si_3N_4 ceramics, we have used the high-angle (HA) ADF-STEM imaging method combined with aberration correction for the probe formation. In this study, a lanthanum oxide (La_2O_3) doped β - Si_3N_4 sintered sample was selected for characterization. La_2O_3 additions are known to strongly promote anisotropic grain growth, resulting in the highest aspect ratio grains compared with other rare-earth oxides⁵.

Figure 1 shows both HAADF-STEM (Fig. 1a) and bright-field (BF)-STEM (Fig. 1b) images of a typical region centred on an intergranular film, recorded simultaneously. The grain on the right in each image is aligned with the [0001] projection of β - Si_3N_4 , so that the (10 $\bar{1}$ 0) prismatic boundary plane is set at an 'edge on' condition. The grain on the left side is oriented with the (01 $\bar{1}$ 1) planes normal to the image surface. In the HAADF-STEM image, bright spots inside the grains correspond to the Si columns, and the bright vertical band structure at the centre of the image indicates the

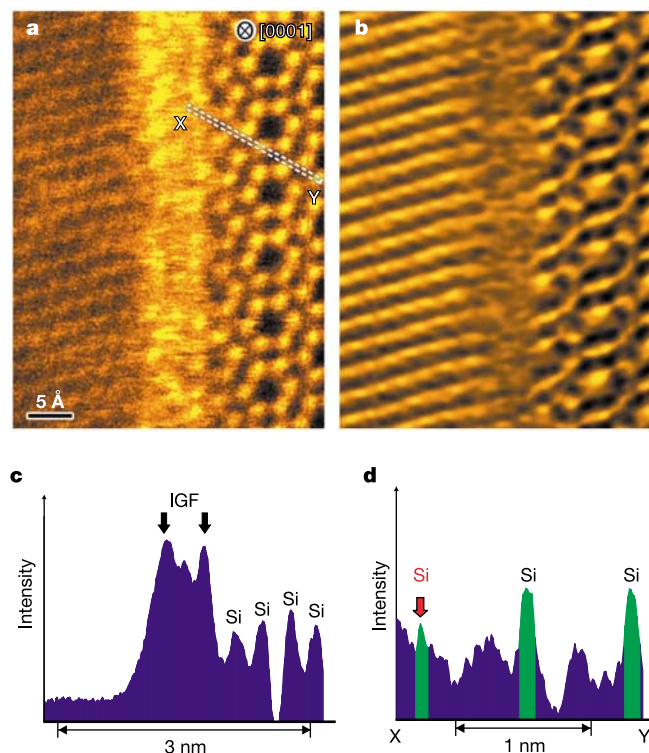


Figure 1 Atomic-resolution scanning transmission electron microscope (STEM) images of an intergranular film (IGF) in La-doped β - Si_3N_4 . **a**, **b**, High-angle annular dark-field (HAADF)-STEM (**a**) and bright-field (BF)-STEM (**b**) images were simultaneously recorded at the same position. The right-hand crystal is aligned with the [0001] projection, and the left-hand crystal is oriented with the (01 $\bar{1}$ 1) planes normal to the image surface in this condition. The bright band seen at the centre of **a** corresponds to the IGF, and brightest spots indicate the location of La atoms. **c**, The image intensity profile across the IGF summed along **a**. The image intensity is higher at the IGF than in the grain interiors, but an intensity variation can be also seen within the film. The interfaces between IGF and grains have much higher intensities than inside the IGF, indicating the preferential segregation of La to the grain surfaces. **d**, The intensity line profile across X–Y, from the edge of the IGF to the grain interior, as defined in **a**. The two green intensity peaks on the right-hand side correspond to the Si columns in the Si_3N_4 lattice. An intensity peak (indicated by the red arrow) is also found that corresponds to the Si column at the edge of the IGF. The intensity here is weaker than that of Si columns in the grain interior, suggesting that the Si column at the edge is not perfect but distorted or partially occupied along the projected direction.

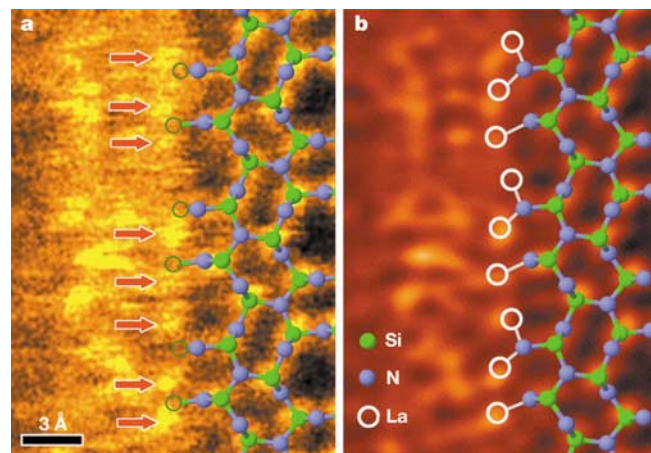


Figure 2 Magnified HAADF-STEM images of the interface between the IGF and the prismatic surface of an β - Si_3N_4 grain. The β - Si_3N_4 lattice structure is superimposed on the images. **a**, La atoms are observed as the bright spots (denoted by red arrows) at the edge of the IGF. The positions of La atoms are shifted from that of Si atoms based on the extension of the β - Si_3N_4 lattice structure; these expected positions are shown by open green circles. **b**, The Pixon (<http://www.pixon.com>) reconstructed image of **a**, showing the La segregation sites more clearly. The predicted La segregation sites obtained by the first-principles calculations are shown by the open white circles. These theoretical predictions agree well with the experimental result. The slight deviations (<1 Å) from some of the theoretically predicted sites probably indicate the influence of atoms in the IGF.

position of the IGF, which is confirmed by the amorphous-like contrast in the BF-STEM image of the same region. The estimated IGF thickness is about 1 nm in both images. The strong bright contrast in the IGF is due to the presence of atoms with high atomic number, in this case, La ($Z = 57$). STEM-electron energy-loss spectroscopy (EELS) spectra also confirmed the presence of La within the intergranular film. Notice that the very strong intensity in the HAADF image is observed at the IGF/grain interfaces and a minimum occurs around the centre of the IGF. Figure 1c shows the image intensity profile across the boundary taken across the width of Fig. 1a. It is clearly seen that the maximum image intensities appear at both IGF/crystalline interfaces, resulting in a bimodal intensity distribution across the IGF. The strong image intensity zone along the right edge of the IGF is located where the terminal Si columns in the right-hand grain would have existed, suggesting that this zone represents the first cation layer attached to the β - Si_3N_4 terminating surface. This is consistent with the prismatic surface of β - Si_3N_4 being terminated by a nitrogen plane, and the cations in the IGF being bonded to it forming the strong image intensity layer.

Figure 2a is a magnified HAADF-STEM image of the IGF/prismatic crystalline interface. The β - Si_3N_4 lattice structure is superimposed on the image. At the strong intensity interfacial zone, La atoms are readily observed as bright spots (denoted by red arrows). Note that the positions of the La atoms are shifted from that of Si atoms based on the extension of the β - Si_3N_4 structure; these expected positions are shown by open green circles. The bright spots within the interfacial zone typically occur at greater distances from nearby N on the terminal grain surface than the expected Si positions, suggesting that La has a larger stable bond length than Si. On the other hand, the expected Si sites also have weaker, but definite, image intensities. The occupancy of these Si sites at the edge of the IGF was confirmed via the intensity profile taken along the white dashed line (X–Y) in Fig. 1a, as shown in Fig. 1d. This line scan runs from the edge of the IGF into the bulk β - Si_3N_4 aligned with the periodic Si columns in the grain. The periodic strong green intensity peaks represent the position of Si columns in the bulk. As indicated by the red arrow, an intensity peak is also found at the very edge of the IGF, near the extended Si position of the lattice. The intensity is weaker than that of Si columns in the bulk crystal, indicating that the column is not perfect, but distorted by the amorphous film or only partially occupied by Si along the projected direction. The same tendency is observed everywhere along the interface. This indicates that the strong image intensity zone at the edge of the IGF is formed by both La and Si atoms.

First-principles atomic cluster calculations¹⁷ were carried out to determine the details of the stable adsorption sites and the energies of La versus Si using a N-terminated Si_3N_4 surface cluster model. To predict the stable La sites, prismatic surface cluster models with La attached were constructed, and then the structures were relaxed to allow the La atoms to reach their stable positions. As information about the precise structure of the amorphous film is lacking, we assumed a bare N-terminated surface model in this study. It was found that there are three independent stable La sites (defined by local energy minima) per surface unit cell along the N-terminated prismatic planes. Calculations show that these La sites have binding energies 0.5–5.3 eV higher than Si at its most stable surface site. Moreover, the La–N bonds are longer than comparable Si–N bonds by about 0.34–0.48 Å. Theoretically predicted stable La sites are superimposed (open white circles) on the Pixon (<http://www.pixon.com>) reconstructed image shown in Fig. 2b. The reconstructed image clearly shows the strong intensity cation sites at the IGF/crystal interface. Despite the assumption of a bare surface, the observed and predicted La positions exhibit remarkable agreement, indicating the validity of the present calculations.

The enhancement of anisotropic grain growth by La_2O_3 additions⁵ must derive from preferential segregation¹⁸ that sup-

presses diametrical growth involving the prismatic surfaces. The present results clearly show that the suppressed growth originates from the preferential segregation (and strong bonding) of La atoms to the grain surfaces; that is, the La atoms partially reside in cation sites normally available for Si at the grain surfaces, effectively hindering the Si attachment necessary for the growth reaction. The present findings provide an atomic level view of where the dopant La atoms prefer to reside in the material, and how they control the microstructure development. These results are consistent with recent first-principles (local density) calculations on the surface segregation of the rare earths competing with Si in Si_3N_4 based ceramics¹⁸. Differential binding energy, which differentiates rare-earth versus Si segregation strengths in environments of variable O/N ratios, was devised to distinguish the elements that remain largely bound within the amorphous phase from those that preferentially segregate to the Si_3N_4 grain surfaces (for example, La). Excellent correlation exists between the calculated differential binding energy values and measured diametric/length grain growth ratios for a variety of rare-earth elements, confirming the importance of the grain growth controlling mechanism induced by the preferential dopant segregation. We believe that the ability to combine atomistic simulations with direct atomic-scale observations of dopant atoms should significantly assist the development of structural ceramic materials. □

Methods

Atomic-resolution images were taken with a dedicated 300-kV (VG Microscopes, HB603U) scanning transmission electron microscope equipped with aberration corrector (Nion Co.), providing a minimum probe diameter of less than 1 Å. The probe convergence semiangle is approximately 22 mrad, and the HAADF image was recorded with detector angle ranges of about 35–300 mrad. The specimen for STEM observations was prepared as follows. The bulk material was mechanically polished to be about 50 µm in thickness using a precision polishing system (Allied High Tech Products, Inc., TechPrep), and then dimpled (Gatan Inc., Model 656) to have a thickness less than 10 µm before ion thinning. In the ion thinning process (Gatan Inc., Model 691), the initial accelerating voltage of Ar ions was 4.5 kV with an incident beam angle of 7°, and gradually decreased to 2 kV as thinning progressed. Finally, a low voltage ion thinning (1 kV) was applied to remove any surface damage layers (Fischione Instruments Inc., Model 1010). Thin amorphous carbon film (<3 nm) was deposited on the sample to avoid charging effects.

Received 24 November 2003; accepted 9 February 2004; doi:10.1038/nature02410.

- Chen, I. W., Becher, P. F., Mitomo, M., Petzow, G. & Yen, T.-S. (eds) *Silicon Nitride Ceramics—Scientific and Technological Advances* (MRS Proceedings, Mater. Res. Soc. 287, Pittsburgh, Pennsylvania, 1993).
- Hoffmann, M. J. & Petzow, G. (eds) *Tailoring of Mechanical Properties of Si_3N_4 Ceramics* (NATO ASI Series E, Applied Sciences Vol. 276, Kluwer Academic, Dordrecht, 1994).
- Lange, F. F. Relation between strength, fracture energy, and microstructure of hot pressed Si_3N_4 . *J. Am. Ceram. Soc.* **56**, 518–522 (1973).
- Tani, E., Umebayashi, K., Kishi, K. & Kobayashi, K. Gas-pressure sintering of Si_3N_4 with concurrent addition of Al_2O_3 and 5wt% rare earth oxide: High fracture toughness Si_3N_4 with fiber-like structure. *Am. Ceram. Soc. Bull.* **65**, 1311–1315 (1986).
- Hoffmann, M. J., Gu, H. & Cannon, R. M. in *Interfacial Engineering for Optimized Properties II* (eds Carter, C. B., Hall, E. L., Briant, C. L. & Nutt, S.) 65–74 (MRS Proceedings 586, Mater. Res. Soc., Warrendale, Pennsylvania, 2000).
- Becher, P. F. *et al.* Microstructural design of silicon nitride with improved fracture toughness: I, Effects of grain shape and size. *J. Am. Ceram. Soc.* **81**, 2821–2830 (1998).
- Sun, E. Y. *et al.* Microstructural design of silicon nitride with improved fracture toughness: II, Effects of yttria and alumina additives. *J. Am. Ceram. Soc.* **81**, 2831–2840 (1998).
- Sun, E. Y. *et al.* Debonding behavior between β - Si_3N_4 whiskers and oxynitride glasses with or without an epitaxial β - SiAlON interface layer. *Acta Mater.* **47**, 2777–2785 (1999).
- Krämer, M., Hoffmann, M. J. & Petzow, G. Grain growth studies of silicon nitride dispersed in an oxynitride glass. *J. Am. Ceram. Soc.* **76**, 2778–2784 (1993).
- Kleebe, H.-J. Structure and chemistry of interfaces in Si_3N_4 ceramics studied by transmission electron microscopy. *J. Ceram. Soc. Jpn* **105**, 453–475 (1997).
- Gu, H., Pan, X., Cannon, R. M. & Rühle, M. Dopant distribution in grain-boundary films in calcia-doped silicon nitride ceramics. *J. Am. Ceram. Soc.* **81**, 3125–3135 (1998).
- Ziegler, A., Kisielowski, C., Hoffmann, M. J. & Ritchie, R. O. Atomic resolution transmission electron microscopy of the intergranular structure of a Y_2O_3 -containing silicon nitride ceramic. *J. Am. Ceram. Soc.* **86**, 1777–1785 (2003).
- Batson, P. E., Dellby, N. & Krivanek, O. L. Sub-ångström resolution using aberration corrected electron optics. *Nature* **418**, 617–620 (2002).
- Krivanek, O. L., Nellist, P. D., Dellby, N., Muffitt, M. F. & Szilagy, Z. Towards sub-0.5 angstrom electron beams. *Ultramicroscopy* **96**, 229–237 (2003).
- Pennycook, S. J. & Jesson, D. E. High-resolution Z-contrast imaging of crystals. *Ultramicroscopy* **37**, 14–38 (1991).
- Pennycook, S. J. Structure determination through Z-contrast microscopy. *Adv. Imag. Electron Phys.* **123**, 173–206 (2002).

17. Averill, F. W. & Painter, G. S. Symmetrized partial-wave method for density functional cluster calculations. *Phys. Rev. B* **50**, 7262–7267 (1994).
18. Painter, G. S., Becher, P. F., Shelton, W. A., Satet, R. L. & Hoffmann, M. J. Differential binding energies: effects of rare-earth additions on grain growth of β - Si_3N_4 and ceramic microstructure. *Phys. Rev. Lett.* submitted.

Acknowledgements We thank R. L. Satet and M. J. Hoffmann for supplying the silicon nitride ceramics used in this study. This work was supported by the US Department of Energy, Office of Basic Energy Sciences, Division of Materials Sciences and Engineering under contract with UT-Battelle, LLC. N.S. is a fellow of the Japan Society for the Promotion of Science (JSPS).

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to N.S. (shibatan@ornl.gov).

Predictability of El Niño over the past 148 years

Dake Chen^{1,2}, Mark A. Cane¹, Alexey Kaplan¹, Stephen E. Zebiak¹ & Daji Huang²

¹Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York 10964, USA

²Laboratory of Ocean Dynamic Processes and Satellite Oceanography, State Oceanic Administration, Hangzhou, China

Forecasts of El Niño climate events are routinely provided and distributed, but the limits of El Niño predictability are still the subject of debate. Some recent studies suggest that the predictability is largely limited by the effects of high-frequency atmospheric ‘noise’^{1–7}, whereas others emphasize limitations arising from the growth of initial errors in model simulations^{8–10}. Here we present retrospective forecasts of the interannual climate fluctuations in the tropical Pacific Ocean for the period 1857 to 2003, using a coupled ocean–atmosphere model. The model successfully predicts all prominent El Niño events within this period at lead times of up to two years. Our analysis suggests that the evolution of El Niño is controlled to a larger degree by

self-sustaining internal dynamics than by stochastic forcing. Model-based prediction of El Niño therefore depends more on the initial conditions than on unpredictable atmospheric noise. We conclude that throughout the past century, El Niño has been more predictable than previously envisaged.

Present estimates of El Niño’s predictability are mostly based on retrospective predictions for the last two or three decades, encompassing a relatively small number of events^{8–11}. With so few degrees of freedom, the statistical significance of such estimates is questionable. In principle, predictability can also be estimated by perturbing initial conditions in numerical model experiments, but the answer is model dependent, and existing models have not been shown to be realistic enough for this purpose. El Niño is evident in instrumental observations dating back to the mid-nineteenth century and in proxy data sets over much longer periods, but no successful attempt to ‘hindcast’ the historic El Niño events before the mid-twentieth century has been reported. This is due partly to the lack of adequate data for model initialization and partly to the inability of present models to make effective use of available data. The study reported here represents the first (to our knowledge) retrospective forecast experiment spanning the past one-and-a-half centuries, using only reconstructed sea surface temperature (SST) data¹² for model initialization.

The intrinsic predictability of El Niño is surely limited, but there has been considerable debate about what the limitations really are^{1,13,14}. Classic theories consider El Niño and the Southern Oscillation (ENSO) as a self-sustaining interannual fluctuation in the tropical Pacific^{15,16}, being chaotic yet deterministic^{17,18}. Thus its predictability is largely limited by the growth of initial errors, and the potential forecast lead time is likely to be of the order of years^{8,10,19}. On the other hand, some recent studies emphasize the importance of atmospheric noise^{2–4}, particularly the so-called westerly wind bursts in the western equatorial Pacific^{5–7}. In such a scenario, ENSO is a damped oscillation sustained by stochastic forcing, and its predictability is more limited by noise than by initial errors. This implies that most El Niño events are essentially unpredictable at long lead times, because their development is often accompanied by high-frequency forcing. Such a view is not supported by the present findings.

The observed and predicted SST anomalies averaged in the central equatorial Pacific are shown in Fig. 1a. (See Methods section

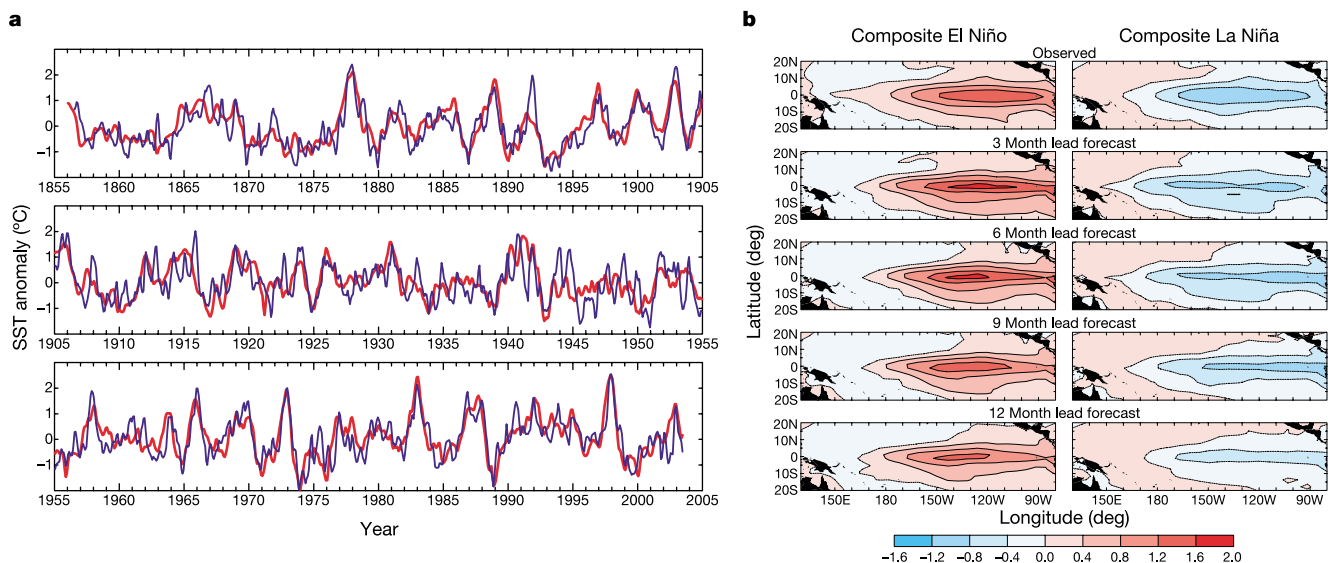


Figure 1 Retrospective predictions of El Niño and La Niña in the past 148 yr. **a**, Time series of SST anomalies averaged in the NINO3.4 region (5° S–5° N, 120–170° W). The red curve is monthly analysis of ref. 12 and the blue curve is the LDE05 prediction at

6-month lead. **b**, Composite El Niño and La Niña from 24 warm events and 23 cold events. Top panels are observations, and the rest are predictions at different lead times. The colour bar shows the range of SST anomalies in degrees Celsius.

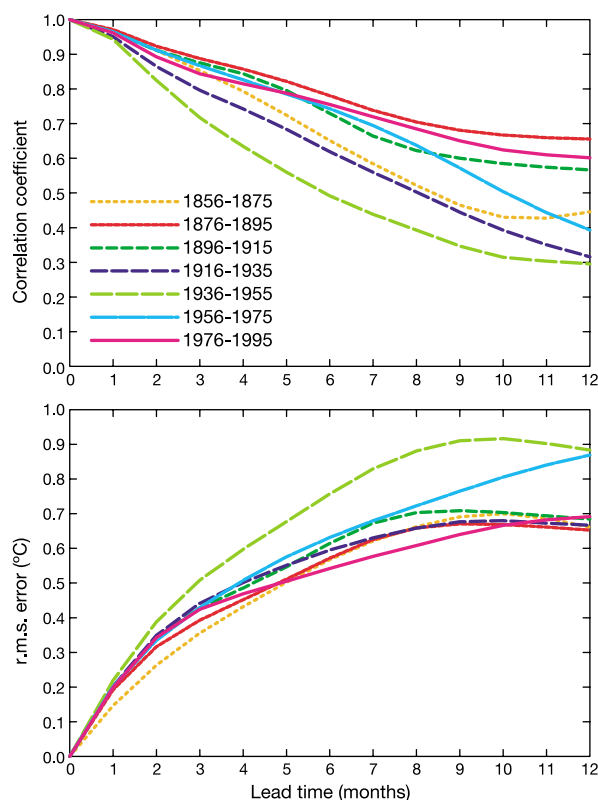


Figure 2 Anomaly correlations and r.m.s. errors between the observed and the predicted values of the NINO3.4 index. These are shown as a function of lead time, for seven consecutive 20-yr periods since 1856.

for a description of the model and data used in this study.) At a 6 month lead, the model was able to predict most of the warm and cold events that occurred in the past 148 yr, especially the relatively large El Niños and La Niñas. The characteristics of the interannual variability obviously have changed with time. Although the strong and regular oscillations in the late twentieth century resemble what happened about a century earlier, there were relatively quiet periods without much activity. If we define an El Niño as a warm event when the NINO3.4 anomaly index is greater than 1 °C, and a La Niña as a cold event of the same amplitude, then there have been 24 of the former and 23 of the latter since 1856. Composites of these observed El Niños and La Niñas are displayed in Fig. 1b, together with corresponding forecast composites. On average, the spatial patterns of both El Niño and La Niña are well captured by the model.

ENSO's predictability depends on the time period from which it is estimated^{9,20,21}. This is evident in Fig. 2. For the seven sub-periods of 20 yr each, both anomaly correlations and r.m.s. errors vary over significant ranges, especially at longer lead times. The periods with the highest overall scores, 1876–95 and 1976–95, are dominated by strong and regular ENSO events. Whereas the high scores for the 1976–95 period may not be surprising because the model was trained using data from part of this period, the even higher scores for the 1876–95 period, which is free of artificial skill, indicate that the large El Niños and La Niñas are highly predictable, even with a simple model initialized with only SST data. The lower skill in other periods is a result of there being fewer and smaller events to predict. For instance, during the 1936–55 period, when the predictability was the lowest by both measures, there were no El Niños except for a prolonged warm event in 1940–42.

Figure 3 shows long-lead forecasts for six of the largest warm episodes (as measured by peak NINO3.4 SST) in the past 148 yr. In all cases, the model was able to predict the observed strong El Niños two years in advance, though some errors exist in the forecasted

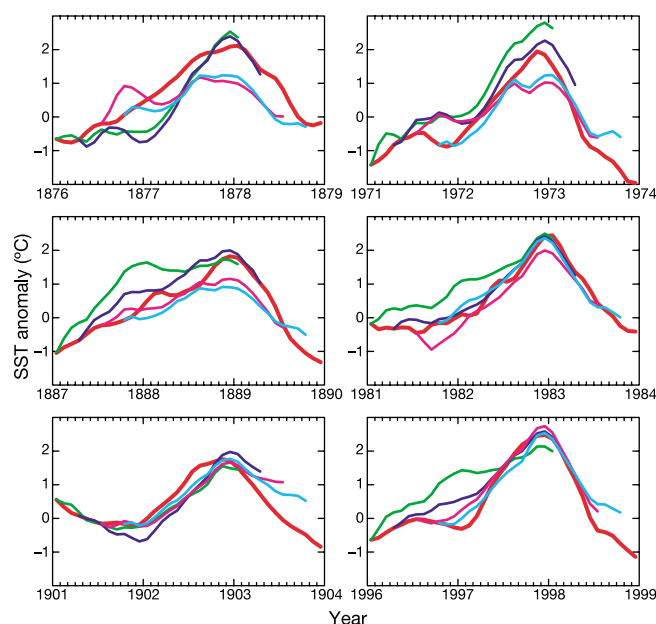


Figure 3 Six of the largest El Niños since 1856. The thick red curves are observed NINO3.4 SST anomalies, and the thin curves of green, blue, magenta and cyan are predictions started respectively 24, 21, 18 and 15 months before the peak of each El Niño.

onset and magnitude of these events. The implication is that the evolution of major ENSO events is largely determined by oceanic initial conditions, and that the effect of subsequent atmospheric noise is generally secondary. It is interesting to note that the model predicts the strong El Niño events in the late nineteenth century, which are notorious for their global impact. These events have been implicated²² in the deaths of tens of millions of people in India, China, Ethiopia, Northeast Brazil and elsewhere. (The disastrous failure of the Indian monsoon in 1877 prompted the establishment of the Observatory in India, later the venue for the work of Walker that forms the foundation of modern understanding of ENSO.) The predictions shown here are, to our knowledge, the first successful retrospective forecasts of these significant historic events.

It has long been recognized that there is a so-called spring barrier in ENSO prediction, a drop of skill in persistence and in all model forecasts across the boreal spring. The skill of our model is also somewhat season-dependent (Fig. 4), as indicated by the relatively low correlations at spring verification time (the straight lines in the figure), but this spring barrier in model prediction is not as severe as that in persistence or in most other forecast models²³. When the model predictions are verified against the observed anomalies exceeding ± 0.7 °C, the skill is much higher and the model can predict across two spring barriers for most start months. It seems that the model does well as long as there is a substantial signal to predict, which is consistent with the results of Figs 2 and 3. There are also some indications of an 'autumn barrier' for lead times longer than one year. This is probably due to the seasonal dependence of the discharge/recharge cycle of the equatorial heat content (or sea level), which leads the SST anomaly by 6–8 months and plays a crucial role in ENSO^{24,25,26}.

In summary, we have performed a retrospective ENSO forecast experiment for the past one-and-a-half centuries, which is several times longer than any previous experiments of this kind. Although the model was trained with recent data, it showed high skill in predicting the historic El Niño events back to the nineteenth century. Most importantly, we demonstrated that the large El Niños are predictable at long lead times. As our model has a self-sustaining internal oscillation and it does not invoke any stochastic

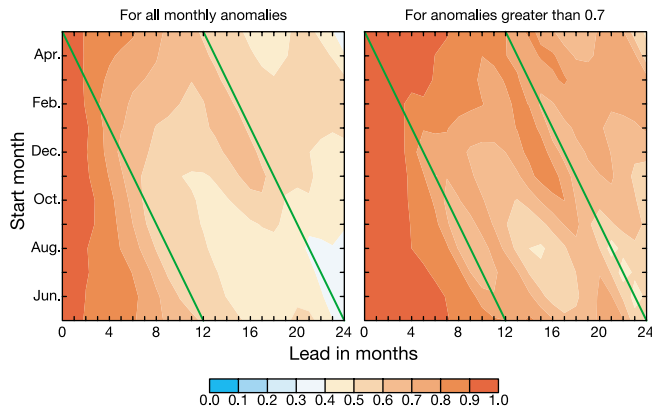


Figure 4 Correlations between observed and predicted NINO3.4 SST anomalies for the 1856–2003 period. These are shown as a function of start month and lead. The straight green lines denote the verification month of May. The left panel is based on all monthly anomalies, while the right panel is for anomalies with amplitudes greater than 0.7 °C. The colour bar shows the range of correlation coefficients.

forcing, this suggests that predictions depend more on initial conditions that determine the phase of ENSO, than on unpredictable atmospheric noise. Although westerly wind bursts do affect the exact onset time and perhaps the amplitude of El Niño, the gross features of ENSO seem to be coded in the large-scale dynamic state. Our results favour the interpretation that the enhanced wind burst activity in the boreal spring preceding large El Niño events is a consequence of those ongoing events²⁷ rather than a cause¹. A practical consequence of our results is a more optimistic view of the possibility of skilful long-lead forecasts of El Niño.

One might suspect that this model does so well in predicting big El Niños because it always wants to predict such events, which means that it might produce as many false alarms as good predictions. To address this issue, we calculated NINO3.4 anomaly correlations for all predicted anomalies exceeding ± 0.7 °C in the 1856–2003 period, and they are 0.58, 0.58 and 0.53 at 18, 21 and 24 month lead times, respectively. These values are about 0.08 smaller than those calculated for observed anomalies exceeding ± 0.7 °C (right panel of Fig. 4), still far above the 95% significance level. This is consistent with the notion that although the model does not work as well during relatively quiet periods, model predictions are not severely plagued by false alarms.

The success of our experiment validates the reliability of the SST data set used here¹². (We obtained similar results with several other newly reconstructed SST data sets^{28,29}.) However, because SSTs are the only data used for model initialization, and because our model is highly simplified and far from perfect, the predictive skill demonstrated here is a lower bound on El Niño's predictability—there is surely room for improvement. As mentioned above, the model's most notable difficulty is in predicting small events and occasions when no events occurred. The key to overcoming this difficulty is to design a data assimilation procedure that can pick the subtle information most relevant to ENSO out of a noisy background. □

Methods

The model used in this study, called LDEO5, is the latest version of an intermediate ocean–atmosphere coupled model^{15,16} widely applied to ENSO investigation and prediction. It differs from its predecessor LDEO4³⁰ in its improved ability to assimilate SST data, which is crucial here as only reconstructed SST data sets are available for such a long-term experiment. In LDEO5, an assimilated SST field not only directly affects the surface wind field as in LDEO4, but also has a persistent effect on the coupled system. The improvement was achieved by including a bias correction term in the model SST equation that statistically corrects for model deficiencies in parameterizing subsurface temperature and surface heat fluxes. The correction was estimated inversely by fitting model SST tendency to observation using data from 1980–2000, and a regression relating this term to the multivariate model state was obtained in the space of empirical orthogonal functions,

using the method of ref. 30. Based on this regression, an interactive correction of SST was then implemented in the model.

The internal variability of LDEO5 is similar to that of LDEO4³⁰; it generates a self-sustaining oscillation with periods of 3–5 yr and amplitudes close to those of observed El Niños. However, the new version has a higher predictive skill when multiple data sets—sea level, winds, SST—are used for initialization, and its skill decreases only slightly when assimilating only SST data. We have to rely on SST data here because tropical Pacific sea level observations are virtually non-existent before 1970, and historic wind information is sparse and poorly calibrated. Note that in the coupled initialization procedure of the LDEO forecast system, assimilated SST data are not simply putting a constraint on the ocean model with SST observations; they translate to surface wind field and subsurface ocean memory. The SST data set used in this study is the reconstructed analysis of ref. 12 for the extended period of 1856–2003. Initialized with this monthly analysis, a forecast with lead times up to 24 months was made from each month of the 148-yr period. The same data set was also used to verify the model predictions.

Received 31 October 2003; accepted 26 February 2004; doi:10.1038/nature02439.

1. Fedorov, A. V., Harper, S. L., Philander, S. G., Winter, B. & Wittenberg, A. How predictable is El Niño? *Bull. Am. Meteorol. Soc.* **84**, 911–919 (2003).
2. Penland, C. & Sardeshmukh, P. D. The optimal growth of tropical sea surface temperature anomalies. *J. Clim.* **8**, 1999–2024 (1995).
3. Moore, A. M. & Kleeman, R. Stochastic forcing of ENSO by the intraseasonal oscillation. *J. Clim.* **12**, 1199–1220 (1999).
4. Thompson, C. J. & Battisti, D. S. A linear stochastic dynamical model of ENSO. Part I: Model development. *J. Clim.* **13**, 2818–2832 (2000).
5. Perigaud, C. M. & Cassou, C. Importance of oceanic decadal trends and westerly wind bursts for forecasting El Niño. *Geophys. Res. Lett.* **27**, 389–392 (2000).
6. McPhaden, M. J. & Yu, X. Equatorial waves and the 1997/98 El Niño. *Geophys. Res. Lett.* **26**, 2961–2964 (1999).
7. Boulanger, J. P. *et al.* Role of non-linear oceanic processes in the response to westerly wind events: New implications for the 1997 El Niño onset. *Geophys. Res. Lett.* **28**, 1603–1606 (2001).
8. Goswami, B. N. & Shukla, J. Predictability of a coupled ocean-atmosphere model. *J. Clim.* **4**, 3–22 (1991).
9. Chen, D., Zebiak, S. E., Busalacchi, A. J. & Cane, M. A. An improved procedure for El Niño forecasting: implications for predictability. *Science* **269**, 1699–1702 (1995).
10. Xue, Y., Cane, M. A. & Zebiak, S. E. Predictability of a coupled model of ENSO using singular vector analysis. Part I: Optimal growth in seasonal background and ENSO cycles. *Mon. Weath. Rev.* **125**, 2043–2056 (1997).
11. Barnston, A. G., Glantz, M. H. & He, Y. X. Predictive skill of statistical and dynamical climate models in SST forecasts during the 1997/98 El Niño episode and the 1998 La Niña onset. *Bull. Am. Meteorol. Soc.* **80**, 217–243 (1999).
12. Kaplan, A. *et al.* Analysis of global sea surface temperature 1856–1991. *J. Geophys. Res.* **103**, 18567–18589 (1998).
13. Chang, P., Ji, L., Li, H. & Flugel, M. Chaotic dynamics versus stochastic processes in El Niño–Southern Oscillation in coupled ocean-atmosphere models. *Physica D* **98**, 301–320 (1996).
14. Latif, M. *et al.* A review of the predictability and prediction of ENSO. *J. Geophys. Res.* **103**, 14375–14393 (1998).
15. Zebiak, S. E. & Cane, M. A. A model El Niño–Southern Oscillation. *Mon. Weath. Rev.* **115**, 2262–2278 (1987).
16. Cane, M. A., Zebiak, S. E. & Dolan, S. C. Experimental forecasts of El Niño. *Nature* **321**, 827–832 (1986).
17. Jin, F. F., Nedin, J. D. & Ghil, M. El Niño on the devil's staircase — annual subharmonic steps to chaos. *Science* **264**, 70–72 (1994).
18. Tziperman, E., Stone, L., Cane, M. A. & Jarosh, H. El Niño chaos: overlapping of resonances between the seasonal cycle and the Pacific ocean-atmosphere oscillator. *Science* **264**, 72–74 (1994).
19. Zebiak, S. E. On the 30–60 day oscillation and the prediction of El Niño. *J. Clim.* **2**, 1381–1387 (1989).
20. Kirtman, B. P. & Schopf, P. S. Decadal variability in ENSO predictability and prediction. *J. Clim.* **11**, 2804–2822 (1998).
21. Balmaseda, M. A., Davey, M. K. & Anderson, D. L. T. Decadal and seasonal dependence of ENSO prediction skill. *J. Clim.* **8**, 2705–2715 (1995).
22. Davis, M. *Late Victorian Holocausts: El Niño Famines and the Making of the Third World* (Verso Books, New York, 2002).
23. Goddard, L. *et al.* Current approaches to climate prediction. *Int. J. Climatol.* **21**, 1111–1152 (2002).
24. Jin, F. F. An equatorial ocean recharge paradigm for ENSO. Part I: Conceptual model. *J. Atmos. Sci.* **54**, 811–829 (1997).
25. Meinen, C. S. & McPhaden, M. J. Observations of warm water volume changes in the equatorial Pacific and their relationship to El Niño and La Niña. *J. Clim.* **13**, 3551–3559 (2000).
26. McPhaden, M. J. Tropical Pacific Ocean heat content variations and ENSO persistent barriers. *Geophys. Res. Lett.* **30**, doi:10.1029/2003GL018672 (2003).
27. Kessler, W. S. EOF representation of the Madden-Julian Oscillation and its connection with ENSO. *J. Clim.* **14**, 3055–3061 (2001).
28. Smith, T. M. & Reynolds, R. W. Extended reconstruction of global sea surface temperatures based on COADS data (1854–1997). *J. Clim.* **16**, 1495–1510 (2003).
29. Rayner, N. A. *et al.* Globally complete analyses of sea surface temperature, sea ice and night marine air temperature, 1871–2000. *J. Geophys. Res.* **108**, doi:10.1029/2002JD002670 (2003).
30. Chen, D., Cane, M. A., Zebiak, S. E., Canizares, R. & Kaplan, A. Bias correction of an ocean-atmosphere coupled model. *Geophys. Res. Lett.* **27**, 2585–2588 (2000).

Acknowledgements This Letter was written during the sabbatical leave of D.C. at the Laboratory of Ocean Dynamic Processes and Satellite Oceanography, Second Institute of Oceanography, Hangzhou, China. This work was supported by the National Oceanic and Atmospheric Administration, and by the National Aeronautics and Space Administration.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.C. (dchen@deo.columbia.edu).

A lower limit for atmospheric carbon dioxide levels 3.2 billion years ago

Angela M. Hessler*, Donald R. Lowe, Robert L. Jones & Dennis K. Bird

Department of Geological and Environmental Sciences, Stanford University, Stanford, California 94305-2115, USA

* Present address: Department of Geology, Grand Valley State University, Allendale, Michigan 49401, USA

The quantification of greenhouse gases present in the Archaean atmosphere is critical for understanding the evolution of atmospheric oxygen, surface temperatures and the conditions for life on early Earth. For instance, it has been argued^{1–4} that small changes in the balance between two potential greenhouse gases, carbon dioxide and methane, may have dictated the feedback cycle involving organic haze production and global cooling. Climate models have focused on carbon dioxide as the greenhouse gas responsible for maintaining above-freezing surface temperatures during a time of low solar luminosity^{5,6}. However, the analysis of 2.75-billion-year (Gyr)-old⁷ palaeosols—soil samples preserved in the geologic record—have recently provided an upper constraint on atmospheric carbon dioxide levels well below that required in most climate models to prevent the Earth's surface from freezing. This finding prompted many to look towards methane as an additional greenhouse gas to satisfy climate models^{1,4,8,9}. Here we use model equilibrium reactions for weathering rinds on 3.2-Gyr-old river gravels to show that the presence of iron-rich carbonate relative to common clay minerals requires a minimum partial pressure of carbon dioxide several times higher than present-day values. Unless actual carbon dioxide levels were considerably greater than this, climate models^{5,6,8} predict that additional greenhouse gases would still need to have a role in maintaining above-freezing surface temperatures.

Siliciclastic sedimentary rocks of the 3.2-Gyr (refs 10, 11) Moodies Group, Barberton Greenstone Belt (BGB), South Africa, include the oldest known non-marine deposits on Earth^{11,12}. Basal clast-supported conglomerates of the Moodies Group have been interpreted to be alluvial fan and braided fluvial deposits on the basis of evidence for limited sediment reworking, periodic subaerial exposure, and unidirectional palaeocurrents¹². The conglomerate is approximately 300 m thick and is dominated by pebbles and cobbles of chert, derived by erosion of deeper-level rocks of the Onverwacht Group, and porphyritic felsic volcanic and volcanoclastic rock, derived mainly from the immediately underlying Fig Tree Group. It also includes less-common (<7%) silicified plutonic rock, silicified ultramafic rock, greywacke, and quartzo-feldspathic sandstone.

Six pebbles showing distinctive dark outer layers or rinds were identified in drill core (73–113 m depth, well below the zone of modern weathering) of the basal Moodies conglomerate, at the Royal Sheba gold mine in the northern BGB (25° 43' S, 31° 10' E). These pebbles are aphanitic or microphyric with a fine-grained matrix of intergrown potassium feldspar, quartz and muscovite,

with traces of carbonate (<3%). In contrast, pebble rinds are characterized by an abundance of quartz and Fe(II)-rich carbonate (>6%, $\text{Fe}_{0.53}\text{Mg}_{0.44}\text{CO}_3$), with lesser amounts of potassium feldspar and muscovite than observed in the pebble cores. The pebble rinds described here are marked by mineralogical changes to the outer 0.25–8 mm of the pebble, and should not be confused with caliche-like carbonate coatings. The Moodies Group conglomerate has experienced regional post-depositional alteration (primarily sericite and carbonate), and we cannot preclude the possibility that the carbonate in the rinds did not undergo ion exchange with diagenetic fluids. In fact, the high Mg^{2+} content at present observed in the rind carbonate is not probably primary, instead indicating partial Mg^{2+} – Fe^{2+} exchange with increasing burial temperatures¹³. However, the modal abundance of quartz and carbonate in the rinds compared with pebble interiors, and their textural fabric (see below), suggest a primary weathering origin for the rinds.

Textural and paragenetic relations demonstrate that the altered pebbles were rounded by river abrasion before the Fe(II)-rich carbonate rinds developed. Two of the pebbles show evidence of fracture during fluvial transport, and the alteration rinds are truncated against the fracture surfaces (Fig. 1). Petrographic and electron microprobe analyses show that the relative abundance of minerals near the fracture surfaces is like that of the pebble cores rather than the rinds¹⁴. If the rinds had formed during a pervasive post-depositional alteration, they would have developed along these fracture surfaces and around other clasts of similar composition. However, alteration rinds are rare on pebbles in this conglomerate, with most clasts of felsic volcanic rock (which make up 51% of the clast population) lacking visible rinds or geochemical evidence of surface alteration. Their rarity in the Moodies Group conglomerate is not surprising; weathering rinds seem to be inherently fragile features of stream-abraded pebbles. Sampling by the authors of modern stream cobbles from the Mojave Desert, the Klamath Mountains and the Sierra Nevada showed that only a small fraction of these cobbles contained the distinct Fe-oxide alteration expected in modern, weathered gravels¹⁴.

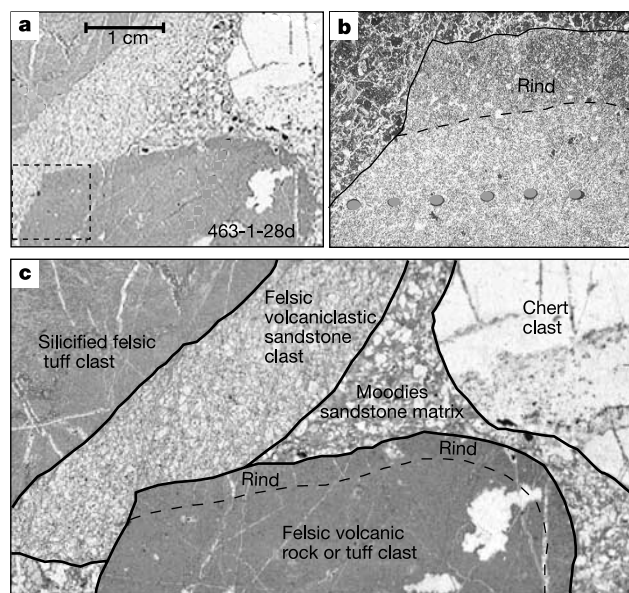


Figure 1 Microscope images of sample pebble 463-1-28d and its weathering rind. **a**, Scanned slide image of pebble 463-1-28d. A dark rind is visible around the outer 2 mm of this clast. **b**, Photomicrograph of the fractured edge of pebble 463-1-28d (see boxed area in **a**). The pebble edge is marked by the solid black line, and the interior extent of the rind is marked by a dotted line. The darkened rind is clearly truncated against the 'fresh' fracture surface. Grey dots were used as guides during microprobe analyses. **c**, A mapped version of **a**, to delineate clast types and matrix surrounding pebble 463-1-28d.

The development of rinds on rounded pebbles, rind truncation along broken pebble surfaces, and the absence of rinds on many surrounding pebbles of similar mineralogy indicate that the rinds formed before final deposition, probably during a stage when the pebbles were 'resting' as part of river-bottom gravels or on gravel bars within the braided stream system. In modern fluvial systems, ground water interacting with channel deposits can provide an environment appropriate for iron-rich carbonate precipitation that is entirely isolated from the ambient atmosphere and/or surface hydrology¹⁵. Such groundwater conditions in fluvial settings arise predominately from the anaerobic subsurface decay of organic material, of which there was an apparent paucity in Archaean fluvial systems. This observation, together with the lack of observed Fe(II)-rich carbonate rinds on post-Archaean braided fluvial or alluvial gravels, drives the assumption that while groundwater may have played a part in weathering the Moodies gravels, it probably did not have a composition far removed from that of surface waters and atmosphere. Ultimately, because there was little organic material to generate isolated environments that form Fe(II)-rich carbonate¹⁵, and because the Fe(II)-rich carbonate rinds survived at least one cycle of erosion and transport in surface waters, their characteristics are useful for constraining surface conditions at the time of Moodies deposition.

The presence of Fe(II)-rich carbonate in the Archaean pebble rinds constrains the oxidation state of the fluvial environment in which they formed. The absence of Fe(II)-rich silicates or Fe(III) oxides in the conglomerate mineralogy suggests that fluid forming the Fe(II)-rich carbonate rinds—whether rain water on a gravel bar, river water, or pore water moving through the river gravels during a temporary resting stage—contained sufficient dissolved Fe(II) to precipitate Fe(II)-rich carbonate upon reaction with the pebbles' mineralogy. Under modern surface conditions, Fe(II) is largely oxidized and immobilized as solid Fe(III) oxides or hydroxides, so for near-surface fluids to transport reduced iron, the logarithm of oxygen fugacity ($\log f_{O_2}$) must be low (at pH 6–7, below approximately -40)¹⁶. Survival of Fe(II)-rich carbonate during transport of fluvial sediment is in accord with abundant evidence for low atmospheric oxygen in the Archaean: the evidence comes from palaeosols^{17–19}, detrital minerals in fluvial systems^{20–23} and mass-

independent fractionation of sulphur isotopes²⁴.

More significantly, the formation of Fe(II)-rich carbonate weathering rinds in the 3.2-Gyr-old Moodies Group conglomerates defines thermodynamic constraints on the lower limit of p_{CO_2} in this near-surface, fluvial, mid-Archaean environment. As a first-order approximation, we model p_{CO_2} constraints using phase relations of Fe(II)-carbonate (siderite) and Fe(II)-layer silicates (greenalite and annite) following the method of ref. 7. Here we consider three end-member divariant mineral assemblages in the system $K_2O-FeO-Al_2O_3-SiO_2-H_2O-CO_2$ as given in Fig. 2: (1) greenalite (model Fe(II) trioctahedral layer silicate)–siderite; (2) annite (model Fe(II) trioctahedral layer silicate)–kaolinite–muscovite (model dioctahedral layer silicate)–siderite; (3) annite–potassium feldspar–siderite. Partial pressures of CO_2 for these assemblages are given in Fig. 2 as a function of temperature at 1 bar total pressure. All three reactions are endothermic with similar enthalpies of reaction, and thus CO_2 increases in a regular fashion with increasing temperature, with a variance in CO_2 of less than an order of magnitude for the three assemblages at any given temperature. Of the three reactions, (2) most closely portrays the mineral paragenesis of the rinds. At 25 °C, the stability of Fe(II)-rich carbonate, siderite and quartz in relation to Fe(II) trioctahedral layer silicate (annite) and kaolinite requires $p_{CO_2} > 2.5 \times 10^{-3}$ bar (Fig. 2). This minimum p_{CO_2} increases with increasing temperatures; at 50 °C and 70 °C, for instance, lower-limit p_{CO_2} increases to 1.4×10^{-2} and 4.7×10^{-2} bar, respectively. If weathering fluids are undersaturated with respect to quartz, reaction lines 1 and 2 in Fig. 2 will be translated towards lower (by at most one order of magnitude) p_{CO_2} values with decreasing activity of SiO_2 , a_{SiO_2} . One should also note that substitution of Mg^{2+} for Fe^{2+} in Fe(II)-rich carbonate will decrease our predicted lower limit of p_{CO_2} . However, equilibrium $Mg^{2+}-Fe^{2+}$ partitioning between the carbonate and Fe(II)-layer silicates, represented by end-member greenalite and

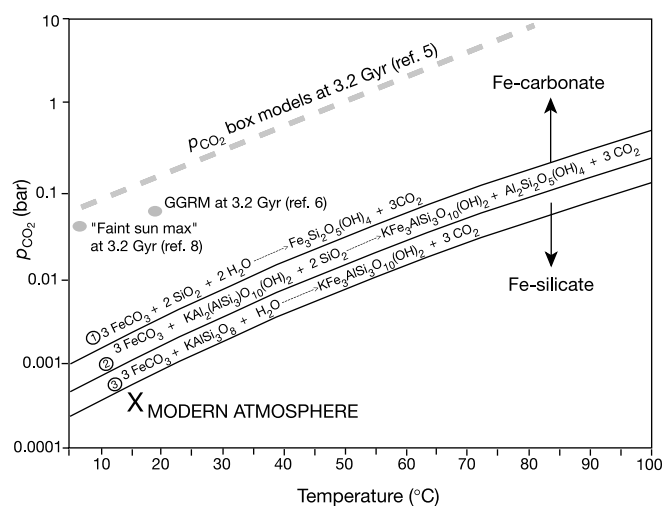


Figure 2 Partial pressure of CO_2 as a function of temperature for the three balanced modelling reactions. Equilibrium constants for each equation at each temperature were derived from SUPCRT92²⁷ with total pressure held constant at 1 bar. Aqueous solution modelling²⁸ shows that the pH of a weathering fluid in contact with the above mineral assemblages is buffered to between 6.5 and 6.8 for these CO_2 fugacity values and temperatures. Values of p_{CO_2} from climate models are included for reference, at their model temperatures. GGRM, 'grey gas radiation modeling'⁶. Modern atmospheric CO_2 ($X_{modern\ atmosphere}$) at 25 °C is 3.5×10^{-2} bar.

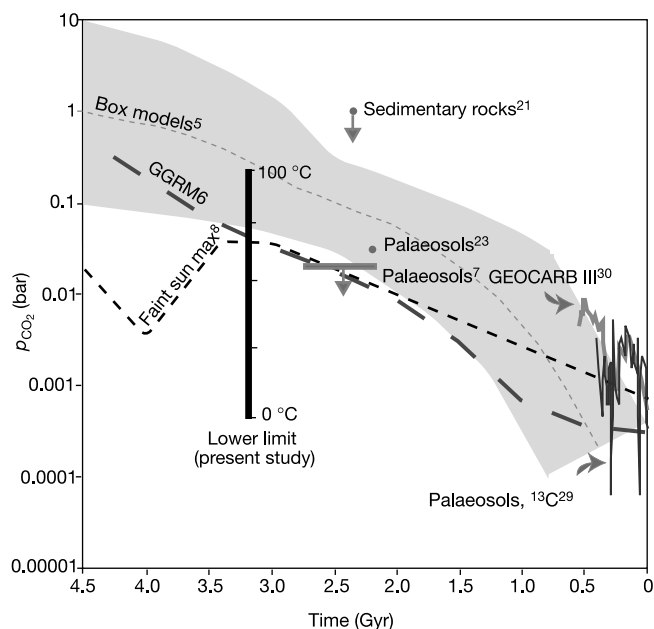


Figure 3 Quantitative estimates of atmospheric CO_2 over geologic time, compiled from both geologic evidence and long-term models. The thin, grey dotted line and the associated grey shaded area represent, respectively, 'best guess' and reasonable range (0–85 °C) for the box model⁵. Estimates of p_{CO_2} from palaeosols⁷ were derived at 15 °C. 'Up' or 'down' arrows indicate minimum or maximum values. Data from this study are indicated by the vertical black bar, and provide modelling results for 0–100 °C. Phanerozoic CO_2 trends based on palaeosol carbon isotopes²⁹ and the GEOCARB III model³⁰ are marked by solid grey and black lines, respectively.

annite, will largely cancel this effect. Therefore curves in Fig. 2 are probably first-order approximations for a wide range of compositions of Fe(II)-rich carbonate phase relations in our model reactions.

Rye and others⁷ established 0.04 bar (an order of magnitude below that required by climate box models³ at 25 °C) as an upper limit for atmospheric p_{CO_2} , based on the presence of Fe(II)-layer silicates and the absence of siderite from lower portions of 2.2–2.75-Gyr palaeosols. Conversely, in this study the presence of Fe(II)-rich carbonate and the absence of Fe(II)-layer silicates in 3.2-Gyr fluvial pebble rinds provide a lower limit on p_{CO_2} of 2.510^{-3} bar (at 25 °C), a value that increases with increasing temperature (Figs 2 and 3). Temperature dependence (0–100 °C) of the minimum p_{CO_2} required for Fe(II)-rich carbonate stability is shown by the scale bar in Fig. 3 at 3.2 Gyr, which can be compared with climate models and geologic observations of p_{CO_2} as a function of geologic time, and as a function of temperature in Fig. 2 at 3.2 Gyr.

Relationships between our model reactions and climate models at 3.2 Gyr (Fig. 2) suggest two contrasting geologic scenarios. If climate-model^{5,6,8} estimates of p_{CO_2} are correct at 3.2 Gyr, our modelling defines an unusual geologic environment, where annite, greenalite and probably other Fe(II) trioctahedral layer silicates were highly unstable. On the other hand, if the Earth's near-surface and weathering environment at 3.2 Gyr was buffered by near-equilibrium between Fe(II)-rich carbonate and common clay minerals, there is approximately two orders of magnitude discrepancy between our p_{CO_2} estimate and that required by climate models^{5,6,8} at temperatures between 0 and 100 °C. At 25 °C, this translates to 0.2 bar p_{CO_2} , and at more extreme temperatures (70 ± 15 °C, as suggested in ref. 25) the equivalent of 6.0 bar CO_2 , thus requiring variable amounts of other greenhouse gases (perhaps methane^{1,26}) to satisfy constraints of global climate models and the stability of clay minerals at 3.2 Gyr (Fig. 2).

The occurrence of Fe(II)-rich carbonate as weathering rinds on pebbles in 3.2-Gyr river gravels of the Barberton Group suggest a higher- CO_2 (and possibly lower- CH_4) atmosphere than at present during the formation of siderite-free 2.2–2.75-Gyr palaeosols investigated in ref. 7. As Earth's global atmospheric, tectonic and biotic regimes evolved during the 0.45 Gyr between 3.2 and 2.75 Gyr ago, geologic evidence suggests an atmospheric decrease in CO_2 and a possible increase in CH_4 or other greenhouse gases (Fig. 3), perhaps following respectively increased silicate weathering on stabilized continents and a growing biogenic source for methane. These shifts probably did not follow the continuous linear trend as indicated by climate models, but might have mimicked the extrema of Phanerozoic concentrations illustrated in Fig. 3. There can be little doubt about the importance of empirical geologic observations for constraining future climate models of Earth's early atmosphere. □

Received 5 May 2003; accepted 9 March 2004; doi:10.1038/nature02471.

- Pavlov, A. A., Kasting, J. F., Brown, L. L., Rages, K. A. & Freedman, R. Greenhouse warming by CH_4 in the atmosphere of early Earth. *J. Geophys. Res.* **105**, 11981–11990 (2000).
- Sagan, C. & Chyba, C. The early faint sun paradox: Organic shielding of ultraviolet-labile greenhouse gases. *Science* **276**, 1217–1221 (1997).
- Brown, L. K. *Numerical Models of Reducing Primitive Atmospheres on Earth and Mars*. Thesis, Penn State Univ. (1999).
- Catling, D. C., Zahnle, K. J. & McKay, C. P. Biogenic methane, hydrogen escape, and the irreversible oxidation of the early Earth. *Science* **293**, 839–843 (2001).
- Kasting, J. F. Earth's early atmosphere. *Science* **259**, 920–926 (1993).
- Hart, M. H. Evolution of the atmosphere of the Earth. *Icarus* **33**, 23–29 (1978).
- Rye, R., Kuo, P. & Holland, H. D. Atmospheric carbon dioxide concentrations before 2.2 billion years ago. *Nature* **378**, 603–605 (1995).
- Sleep, N. H. & Zahnle, K. Carbon dioxide cycling and implications for climate on ancient Earth. *J. Geophys. Res.* **106**, 13733–1399 (2001).
- Rye, R. & Holland, H. D. Life associated with a 2.76 Ga ephemeral pond? Evidence from Mount Roe paleosol. *Geology* **28**, 483–486 (2000).
- Kamo, S. L. & Davis, D. W. Reassessment of Archean crustal development in the Barberton Mountain Land, South Africa, based on U–Pb dating. *Tectonics* **13**, 167–192 (1994).
- Heubeck, C. E. & Lowe, D. R. Depositional and tectonic setting of the Archean Moodies Group, Barberton Greenstone Belt, South Africa. *Precamb. Res.* **68**, 257–290 (1994).
- Eriksson, K. A. Alluvial and destructive beach facies from the Archean Moodies Group, Barberton

- Mountain Land, South Africa and Swaziland. *Mem. Can. Soc. Petrol. Geol.* **5**, 287–311 (1978).
- Mozley, P. S. Relationship between depositional environment and the elemental composition of early diagenetic siderite. *Geology* **17**, 704–706 (1989).
- Hessler, A. M. *Evidence for Climate and Weathering in Siliciclastic Sedimentary Rocks of the 3.2 Ga Moodies Group, Barberton Greenstone Belt, South Africa*. Thesis, Stanford Univ. (2001).
- Postma, D. Pyrite and siderite formation in brackish and freshwater swamp sediments. *Am. J. Sci.* **282**, 1151–1183 (1982).
- Eugster, H. P. & Chou, I. The depositional environments of Precambrian banded iron-formations. *Econ. Geol.* **68**, 1144–1168 (1973).
- Rye, R. & Holland, H. D. Paleosols and the evolution of atmospheric oxygen; a critical review. *Am. J. Sci.* **298**, 621–672 (1998).
- Macfarlane, A. W., Danielson, A. & Holland, H. D. Geology and major trace element chemistry of late Archean weathering profiles in the Fortescue Group, Western Australia: implications for atmospheric P_{O_2} . *Precamb. Res.* **65**, 297–317 (1994).
- Yang, W. & Holland, H. D. The Hekpoort paleosol profile in Strata 1 at Gaborone, Botswana: Soil formation during the great oxidation event. *Am. J. Sci.* **303**, 187–220 (2003).
- Rasmussen, B. & Buick, R. Redox state of the Archean atmosphere; evidence from detrital heavy minerals in ca. 3250–2750 Ma sandstones from the Pilbara Craton, Australia. *Geology* **27**, 115–118 (1999).
- Krupp, R. et al. The early Precambrian atmosphere and hydrosphere: Thermodynamic constraints from mineral deposits. *Econ. Geol.* **89**, 1581–1598 (1994).
- Grandstaff, D. E. Origin of uraniferous conglomerates at Elliot Lake, Canada and Witwatersrand, South Africa; implications for oxygen in the Precambrian atmosphere. *Precamb. Res.* **13**, 1–26 (1980).
- Holland, H. D. *The Chemical Evolution of the Atmosphere and Oceans* (Princeton Univ. Press, New Jersey, 1984).
- Farquhar, J., Bao, H. & Thiemens, M. Atmospheric influence of Earth's earliest sulfur cycle. *Science* **289**, 756–758 (2000).
- Knauth, L. P. & Lowe, D. R. High Archean climatic temperature inferred from oxygen isotope geochemistry of cherts in the 3.5 Ga Swaziland Supergroup, South Africa. *Geol. Soc. Am. Bull.* **115**, 566–580 (2003).
- Kiehl, J. T. & Dickinson, R. E. A study of the radiative effects of enhanced atmospheric CO_2 and CH_4 on early Earth surface temperatures. *J. Geophys. Res.* **92**, 2991–2998 (1987).
- Johnson, J., Oelkers, E. & Helgeson, H. SUPCRT92: a software package for calculating the standard molar thermodynamic properties of minerals, gases, aqueous species, and reactions from 1 to 5000 bar and to 1000 degrees C. *Comput. Geosci.* **18**, 899–947 (1992).
- Wolery, T. J. *EQ3/6: A Software Package for Geochemical Modeling of Aqueous Systems* (The Regents of the University of California and Lawrence Livermore National Laboratory, Berkeley, 1997).
- Ekat, D. D., Cerling, T. E., Montanez, I. P. & Tabor, N. J. A 400 million year carbon isotope record of pedogenic carbonate: Implications for paleoatmospheric carbon dioxide. *Am. J. Sci.* **299**, 805–827 (1999).
- Berner, R. A. & Kothvala, Z. GEOCARB III: A revised model of atmospheric CO_2 over Phanerozoic time. *Am. J. Sci.* **301**, 182–204 (2001).

Acknowledgements We appreciate the help and samples we received from Avgold and ETC Mines, Barberton, South Africa. Electron microprobe analysis was performed at the Center for Materials Research and Department of Geological and Environmental Sciences at Stanford University. We thank T. Fridriksson, K. Lemke, C. Oze and A. Fildani for technical support and advice, and N. Sleep for comments on early versions of this manuscript. This work was supported by the NASA Exobiology Program (D.R.L.) and the National Science Foundation (D.K.B.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.H. (hesslera@gvsu.edu).

Dynamic response of Permian brachiopod communities to long-term environmental change

Thomas D. Olszewski¹ & Douglas H. Erwin^{2,3}

¹Department of Geology and Geophysics, 3115 TAMU, Texas A&M University, College Station, Texas 77843, USA

²Department of Paleobiology MRC-121, Smithsonian Institution, NMNH, Washington DC 20013, USA

³Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA

The fossil record preserves numerous natural experiments that can shed light on the response of ecological communities to environmental change. However, directly observing the community dynamics of extinct organisms is not possible. As an alternative, neutral ecological models^{1–3} suggest that species

abundance distributions reflect dynamical processes like migration, competition, recruitment, and extinction. Live–dead comparisons suggest that such distributions can be faithfully preserved in the rock record⁴. Here we use a maximum-likelihood approach to show that brachiopod (lamp shell) abundance distributions from four temporally distinct ecological landscapes from the Glass Mountains, Texas (of the Permian period), exhibit significant differences. Further, all four are better fitted by zero-sum multinomial distributions, characteristic of Hubbell's neutral model², than by log-normal distributions, as predicted by the traditional ecological null hypothesis⁵. Using the neutral model as a guide, we suggest that sea level fluctuations spanning about 10 Myr altered the degrees of isolation and exchange among local communities within these ecological landscapes. Neither these long-term environmental changes nor higher-frequency sea level fluctuations resulted in wholesale extinction or major innovation within evolutionary lineages.

The fossil brachiopods used in this study come from the Lower and Middle Permian succession of the Glass Mountains, Texas (Cathedral Mountain, Road Canyon, and Word Formations; Fig. 1). All specimens were collected and described by G. A. Cooper and R. E. Grant, who built a data set of exceptional taxonomic refinement and consistency over 30 years (refs 6–11). Our analyses are based on 855,047 specimens (not all identified to species) from 191 localities, including 512 species of 142 genera and are housed at the National Museum of Natural History, Washington DC, USA (Table 1).

Thanks to excellent exposure and interest from the petroleum industry, rocks of the Permian basin have a well-established sequence stratigraphic framework (Fig. 1)^{12,13}. The approximately 10 Myr interval covered by this study includes four third-order depositional sequences (LD10, GP1–2, GP3 and GP4) with average

duration of 2.5 Myr (third-order depositional sequences record sea level cycles of 1–10 Myr in duration). The stratigraphic stacking pattern of these four sequences is influenced by second-order changes in relative sea level (that is, sea level cycles of 10–100 Myr in duration) showing initial rise in the Leonardian series, maximum inundation during the Roadian stage, and relatively slow fall through the rest of the Guadalupian series.

Each third-order sequence is bounded by an unconformity that represents a major regional disruption of the benthic palaeocommunities due to extensive lateral shifts of depositional environments and changes in habitats. Correspondence analysis indicates that the fossil assemblages of each sequence are compositionally distinct at the species level (Supplementary Information). On the basis of ordination results and the unconformities, we have combined all the collections from each sequence, thereby treating each as a distinct ecological landscape. Although individual collections can be confidently assigned to particular sequences, their stratigraphic and environmental positions within each sequence remain poorly constrained.

The four sequences seem to represent comparable brachiopod communities on the basis of their similarity in generic composition despite high species turnover (65 genera range through three or more sequences). Morphological differences between brachiopod species are often subtle and not related to the gross functional morphology of the organisms. In contrast, genus-level differences are more obvious and provide a crude proxy for the diversity of brachiopod modes of life. Therefore, a consistent pool of genera suggests that each sequence preserves a fundamentally similar range of habitats. Most genera that were not recovered from the entire study interval occur in significantly fewer collections than more ubiquitous genera (Supplementary Information). These genera are preferentially found in the more heavily sampled part of the study interval (the lower two sequences, which include a larger number of collections than the upper two).

Comparisons of modern marine mollusc communities with bulk collections of shell material from associated death assemblages indicate that rank abundances of living communities can be preserved with reasonably high fidelity⁴. Brachiopod shells are of similar size and composition to molluscs, and show equivalent degrees of time-averaging¹⁴, so we treat fossil brachiopod abundance distributions as accurate reflections of the corresponding, ancient living communities.

The species abundance distribution of each sequence was fitted with a zero-sum multinomial and a log-normal distribution using the maximum-likelihood method (Fig. 2). These models provide a way of statistically comparing the abundance distributions of the four sequences and shed light on the processes that may have been influencing these communities. We chose the log-normal and zero-sum multinomial distributions because each is associated with a particular model of how communities work and neither requires any additional information on species interactions or geographic distributions (unlike some other approaches¹⁵). These are also the most commonly compared models in the recent literature^{2,5,16}.

The zero-sum multinomial distribution^{16,17} emerges from Hubbell's neutral model of ecological communities. This model assumes that the total number of individuals in a community of ecologically similar species remains constant and that the survival and reproductive fate of any individual organism is independent of interactions with individuals of the same or other species^{2,3}. Whenever an organism in a community dies, it can be replaced by: (1) recruitment of a species from the local community; (2) recruitment of a species from the surrounding metacommunity; or (3) origination of a new species. The basic patterns of species abundance, geographic distribution, and compositional turnover through time can be predicted using J (the size of the local community), m (the probability of replacement by immigration from the metacommunity), and θ

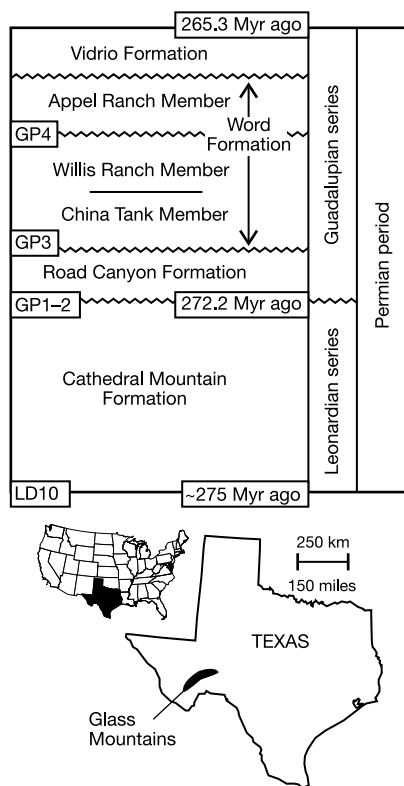


Figure 1 Stratigraphic column and location map. Zigzag lines indicate unconformities bounding depositional sequences¹³. Sequence labels (LD10, GP1–2, GP3 and GP4) are based on the underlying unconformity. Top Guadalupian age from ref. 29. Base Guadalupian age from ref. 30.

Table 1 Sampling and abundance distribution information for each sequence

Sequence	Formation	J	C	S	G	Log-normal			Zero-sum		
						L	μ	σ^2	L	θ	m
LD10	Cathedral Mountain	224,116	81	206	104	570.84	6.83	14.86	556.65	23.06	0.547
GP1–2	Road Canyon	272,614	69	243	107	673.95	7.01	14.93	657.67	27.66	0.826
GP3	lower Word	250,980	25	146	76	411.11	7.79	16.26	401.10	16.84	0.087
GP4	upper Word	45,344	16	97	65	259.22	6.51	12.18	251.99	13.34	0.304

C, number of collections; S, number of species; G, number of genera; μ , mean; σ^2 , variance.

(the fundamental biodiversity number, proportional to the rate of per capita species origination). The lack of interspecific interactions in the neutral model has been criticized as unrealistic^{18,19}. However, the value of the model lies less in its correctness than as a baseline for recognizing non-neutral dynamics when its predicted patterns are violated.

An alternative model of species abundances is the log-normal distribution^{5,20}, which is characterized by fewer rare taxa than the zero-sum multinomial². Calling on the central limit theorem, May has interpreted this distribution as the result of a large number of

geometrically increasing populations that are independently influenced by many identically distributed random variables²¹. However, if there is any possibility of population extinction, the central limit theorem is inapplicable, and abundance distributions are not likely to resemble the log-normal²². Instead, a log-normal-type distribution emerges when a large number of communities from similar types of habitats in which a few species are consistently dominant are added together. The log-normal distribution has also been associated with niche-partitioning models, but these can produce a variety of distributions¹⁵.

The fitted maximum-likelihood distributions for these two models indicate that the zero-sum multinomial provides a better description of brachiopod abundance data than the log-normal distribution for all four sequences (Table 1). The histograms of data from each sequence have a flat-topped appearance, indicating more rare species and more very-common species than expected in the log-normal model (Fig. 2). This is despite the fact that combining samples is expected to produce a log-normal-type curve²². Combining samples from different times during the evolution of a single neutral community (a concern when combining samples within sequences that range over 2.5 Myr), produces abundance distributions with many more rare taxa than even the zero-sum multinomial predicts (Supplementary Information).

In addition to providing a means of comparing the two types of distributions, the maximum-likelihood approach also allows confidence intervals to be placed on the estimated parameters of the zero-sum multinomial model (Fig. 3). These indicate that the abundance distributions of sequences LD10 and GP1–2 differ

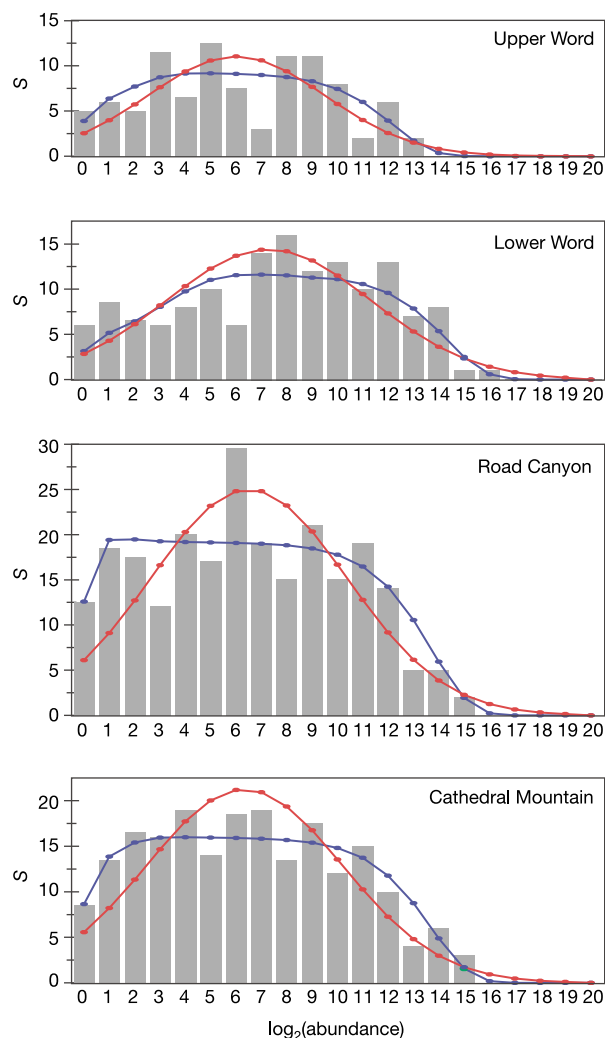


Figure 2 Abundance distributions of species from each sequence. Red line, maximum-likelihood log-normal distribution; blue line, maximum-likelihood zero-sum multinomial distribution; grey bars, observed numbers of species binned into $\log_2(\text{abundance})$ categories²⁰. Species with abundance equal to the boundary between two bins were split in half between the two bins^{15,28}.

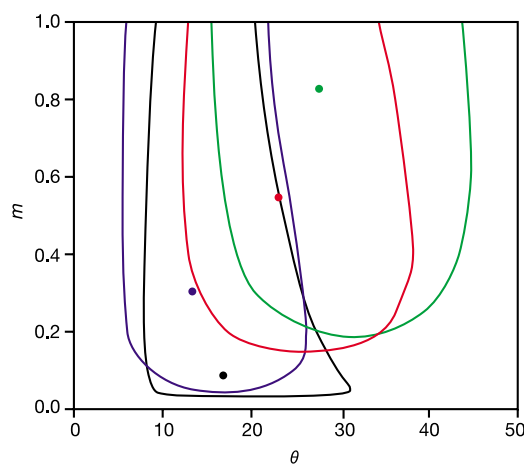


Figure 3 95% confidence boundaries of maximum-likelihood zero-sum multinomial parameters for each sequence. Red, sequence LD10; green, sequence GP1–2; black, sequence GP3; blue, sequence GP4. Data points correspond to the maximum-likelihood estimates of zero-sum multinomial parameters; lines represent the 95% confidence region for each sequence. The maximum-likelihood estimates for sequences GP3 and GP4 fall outside the 95% confidence regions of sequences LD10 and GP1–2. Sequence GP1–2 falls outside the regions of GP3 and GP4, whereas LD10 is just inside the boundaries of GP3 and GP4.

significantly from those of sequences GP3 and GP4 owing to decreases in both θ and m .

Used as a guide to interpreting abundance data, the neutral model suggests that the ecological dynamics of the brachiopod communities in the Glass Mountains were responding to second-order changes in sea level: during sea level rise from sequence LD10 to GP1–2, both θ and m increase to maximum values that coincide with maximum inundation of the basin during sequence GP1–2. As second-order sea level began to drop in sequences GP3 and GP4, both θ and m also decrease relative to their peak values.

Sea level is likely to influence ecological dynamics in at least three ways. First, relative sea level may have correlated with the area available for occupation by brachiopods. A larger area would have increased total population sizes in the Glass Mountains region, allowing otherwise marginally successful rare species to maintain viable populations. Second, the inundated area of the entire Permian basin would have increased as shorelines were pushed landward. This would have increased the opportunity for isolation of local populations in different parts of the basin and the generation of new species. This, in turn, would have led to a potentially more diverse pool of species for recruitment into the local communities of the Glass Mountains region. Third, connections to other marine basins in southwestern North America would have opened or expanded, further enlarging the available species pool.

A good fit by the zero-sum multinomial distribution to the abundances of Glass Mountain brachiopods does not prove that interspecific interactions did not influence the dynamics of these communities. Regardless of the importance of interspecific interactions, however, neutral model parameters seem to correlate with second-order sea level. This suggests that environmental change driven by second-order sea level influenced these Middle Permian ecological landscapes through changes in the degree of isolation and exchange among local communities. Much more rapid perturbations recorded by third-order unconformities did not result in wholesale extinction of, or fundamental innovation within, the evolutionary lineages within these communities. □

Methods

Specimen preparation

All specimens in this study were preserved by silicification, which allowed them to be recovered by dissolving the enclosing carbonate matrix rather than by destructive physical extraction⁶. In addition to preserving the delicate features of brachiopod morphology needed for species-level taxonomy, this method inherently produces bulk samples of all silicified fossil material in a volume of rock.

Stratigraphy

Sequence boundary labels in Fig. 1 are those of ref. 12. The unconformity at the base of the Appel Ranch Member of the Word Formation correlates with the Grayburg–Queen sequence boundary (GP4) in the Guadalupe Mountains²³. The basal unconformity of the San Andres Formation in the Guadalupe Mountains (LD10) correlates with the exposure surface at the base of the Cathedral Mountain Formation in the Glass Mountains²⁴. The boundaries GP1, GP2 and GP3 of ref. 13 are more problematic, but we suggest the following resolution. GP1, which is not well expressed in the Guadalupe Mountains, correlates with the minor exposure surface at the base of the Road Canyon Formation²⁴. GP2, a major surface in the Guadalupe, correlates with the base of the second cycle of the Road Canyon Formation — this is consistent with the conodont transition from *Mesogondolella idahoensis* to *M. nankingensis* that defines the base of the Guadalupian series^{25,26}. Last, GP3, the unconformity at the base of the Grayburg Formation in the Guadalupe Mountains, correlates with the unconformity at the base of the China Tank Member of the Word Formation²⁴. The major faunal turnover between sequences A and B most probably coincides with GP2, but if some of the brachiopod collections come from the lower cycle of the Road Canyon Formation, then the turnover event would coincide with GP1.

Analysis

The maximum-likelihood estimates of the zero-sum multinomial and log-normal functions (Table 1, Figs 2 and 3) were derived by first determining the probability of a species' abundance coming from a $\log_2$ (abundance) category for each distribution (that is, Preston's octaves)²⁰. For the continuous log-normal function, this was simply the difference between the cumulative distribution values at each octave boundary. For the discrete zero-sum multinomial function, the probability of each octave was the sum of all probability values that fell between the boundary values of that bin plus one-half of the probabilities of the abundances at the lower and upper boundary values^{16,17,27}. The data

from each depositional sequence were binned in the same way as the discrete zero-sum multinomial function. Octave binning is necessary because without binning, species with high abundances act as outliers and dominate the likelihood estimate^{25,28}. The negative log-likelihood value for either function equals $L = \sum_{i=1}^n X_i (-\log(P_i))$, where n is the number of octave bins, X_i is the number of species in bin i , P_i is the likelihood of bin i (ref. 28). The 95% confidence regions of the established zero-sum multinomial parameters (Fig. 3) were found by searching for those parameter values that gave a negative log-likelihood that was 2.99 greater than the maximum-likelihood parameters²⁸.

Received 10 December 2003; accepted 5 March 2004; doi:10.1038/nature02464.

- Hughes, R. G. A model of the structure and dynamics of benthic marine invertebrate communities. *Mar. Ecol. Prog. Ser.* **15**, 1–11 (1984).
- Hubbell, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton Univ. Press, Princeton, New Jersey, 2001).
- Bell, G. Neutral macroecology. *Science* **293**, 2413–2418 (2001).
- Kidwell, S. M. Preservation of species abundance in marine death assemblages. *Science* **294**, 1091–1094 (2001).
- McGill, B. J. A test of the unified neutral theory of biodiversity. *Nature* **422**, 881–885 (2003).
- Cooper, G. A. & Grant, R. E. *Permian Brachiopods of West Texas I*–231 (Smithsonian Contributions to Paleobiology 14, Smithsonian Institution Press, Washington DC, 1972).
- Cooper, G. A. & Grant, R. E. *Permian Brachiopods of West Texas, II*, 233–793 (Smithsonian Contributions to Paleobiology 15, Smithsonian Institution Press, Washington DC, 1974).
- Cooper, G. A. & Grant, R. E. *Permian Brachiopods of West Texas, III*, 795–1921 (Smithsonian Contributions to Paleobiology 19, Smithsonian Institution Press, Washington DC, 1975).
- Cooper, G. A. & Grant, R. E. *Permian Brachiopods of West Texas, IV*, 1923–2607 (Smithsonian Contributions to Paleobiology 21, Smithsonian Institution Press, Washington DC, 1976).
- Cooper, G. A. & Grant, R. E. *Permian Brachiopods of West Texas, V*, 2609–3159 (Smithsonian Contributions to Paleobiology 24, Smithsonian Institution Press, Washington DC, 1976).
- Cooper, G. A. & Grant, R. E. *Permian Brachiopods of West Texas, VI*, 3161–3370 (Smithsonian Contributions to Paleobiology 32, Smithsonian Institution Press, Washington DC, 1977).
- Sarg, J. F. & Lehmann, P. J. in *Lower and Middle Guadalupian Facies, Stratigraphy, and Reservoir Geometries, San Andres/Grayburg Formations, Guadalupe Mountains, New Mexico and Texas* (eds Moore, G. & Wilde, G.) 1–8 (SEPM Permian Basin Section, Tulsa, Oklahoma, 1986).
- Sarg, J. F., Markello, J. R. & Weber, L. J. in *Advances in Carbonate Sequence Stratigraphy: Applications to Reservoirs, Outcrops, and Models* (eds Harris, P. M., Saller, A. H. & Simo, J. A.) 11–34 (Society for Sedimentary Geology (SEPM), Tulsa, Oklahoma, 1999).
- Carroll, M., Kowalewski, M., Simoes, M. G. & Goodfriend, G. A. Quantitative estimates of time-averaging in terebratulid brachiopod shell accumulations from a modern tropical shelf. *Paleobiology* **29**, 381–402 (2003).
- Sugihara, G., Bersier, L.-F., Southwood, T. R. E., Pimm, S. L. & May, R. M. Predicted correspondence between species abundances and dendrograms of niche similarities. *Proc. Natl Acad. Sci. USA* **100**, 5246–5251 (2003).
- Volkov, I., Banavar, J. R., Hubbell, S. P. & Maritan, A. Neutral theory and relative species abundance in ecology. *Nature* **424**, 1035–1037 (2003).
- McKane, A. J., Alonso, D. & Sole, R. V. Analytic solution of Hubbell's model of local community dynamics. *Theor. Pop. Biol.* **65**, 67–73 (2004).
- Clark, J. S. & MacLachlan, J. S. Stability of forest biodiversity. *Nature* **423**, 635–638 (2003).
- Fargione, J., Brown, C. S. & Tilman, D. Community assembly and invasion: An experimental test of neutral versus niche processes. *Proc. Natl Acad. Sci. USA* **100**, 8916–8920 (2003).
- Preston, F. W. The commonness and rarity of species. *Ecology* **29**, 254–283 (1948).
- May, R. M. in *Ecology and Evolution of Communities* (eds Cody, M. L. & Diamond, J. M.) 81–120 (Belknap, Cambridge, Massachusetts, 1975).
- Routledge, R. D. The form of species-abundance distributions. *J. Theor. Biol.* **82**, 547–558 (1980).
- Rohr, D. M. et al. in *The Guadalupian Symposium* (eds Wardlaw, B. R., Grant, R. E. & Rohr, D. M.) 5–36 (Smithsonian Contributions to the Earth Sciences 32, Smithsonian Institution Press, Washington DC, 2000).
- Wardlaw, B. R., Ross, C. A. & Grant, R. E. in *The Guadalupian Symposium* (eds Wardlaw, B. R., Grant, R. E. & Rohr, D. M.) 121–126 (Smithsonian Contributions to the Earth Sciences 32, Washington DC, 2000).
- Wardlaw, B. R. in *The Guadalupian Symposium* (eds Wardlaw, B. R., Grant, R. E. & Rohr, D. M.) 37–87 (Smithsonian Contributions to the Earth Sciences 32, Smithsonian Institution Press, Washington DC, 2000).
- Harris, M. T., Lehmann, D. J. & Lambert, L. L. in *The Guadalupian Symposium* (eds Wardlaw, B. R., Grant, R. E. & Rohr, D. M.) 153–173 (Smithsonian Contributions to the Earth Sciences 32, Smithsonian Institution Press, Washington, DC, 2000).
- Nee, S., Harvey, P. H. & May, R. M. Lifting the veil on abundance patterns. *Proc. R. Soc. Lond. B* **243**, 161–163 (1991).
- Hilborn, R. & Mangel, M. *The Ecological Detective: Confronting Models with Data* (Princeton Univ. Press, Princeton, New Jersey, 1997).
- Bowring, S. A. et al. U/Pb zircon geochronology and tempo of the end-Permian mass extinction. *Science* **280**, 1039–1045 (1998).
- Jin, Y., Wardlaw, B. R., Glenister, B. F. & Kotlyar, C. V. Permian chronostratigraphic subdivisions. *Episodes* **20**, 10–15 (1997).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. I. Miller and M. Foote for comments. This research was supported by the Walcott Fund of the Department of Paleobiology of the Smithsonian Institution and the NASA Astrobiology Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.D.O. (tomo@geo.tamu.edu).

Variable female preferences drive complex male displays

Seth W. Coleman¹, Gail L. Patricelli^{1*} & Gerald Borgia^{1,2}

¹Behavior, Ecology, Evolution and Systematics Program, and ²Department of Biology, University of Maryland, College Park, Maryland 20742, USA

* Present address: Macaulay Library, Cornell Laboratory of Ornithology, Ithaca, New York 14850, USA

Complexity in male sexual displays is widely appreciated¹ but diversity in female mate choice has received little attention. Males of many species have sexual displays composed of multiple display traits, and females are thought to use these different traits in mate choice¹. Models of multiple display trait evolution suggest that these traits provide females with different kinds of information in different stages of the mate choice process², or function as redundant signals to improve the accuracy of mate assessment^{3,4}. We suggest that complex male displays might also arise because of variation in female preferences for particular male display traits. The causes of female preference variation have received little attention^{3–7}, and the role of preference variation in shaping complex male displays is unclear. Here we show that in satin bowerbirds (*Ptilonorhynchus violaceus*) female mate choice is a multistage process, where females of different ages use different male display traits in successive stages. Age- and stage-specific female preferences may contribute to explaining the widespread occurrence of multifaceted male displays.

Male satin bowerbirds build specialized stick structures, called bowers, where courtship and copulation take place⁸. Males have multiple display elements combining intense behavioural displays with bower decoration displays^{8,9}. Intense behavioural displays can be both attractive and threatening to females; however, this threat is reduced when males adjust their display in response to female signals that indicate the level of intensity the female will tolerate⁹. Females differ in their tolerance of intense display, with age/experience as a likely cause¹⁰. If young, inexperienced females have a lower tolerance for intense display, then we expect them to emphasize non-threatening display elements, such as bower decorations, in their mate choice decisions. If older, more experienced females are both more tolerant of intense display and are better able to assess males using intense display, then we expect them to emphasize male display intensity in their mate choice decisions. Here we consider two questions: (1) do females show age-related differences in how they use male display traits to choose mates, and (2) are those differences related to female tolerance of male display intensity? To test these hypotheses, we augmented the blue decorations (Fig. 1; see also Supplementary Fig. S1) at a subset of bowers in a natural population of satin bowerbirds, separated females into age classes (first-year females, second-year females, three-plus females (females with two or more years of mating experience)), and then monitored individual female's mating-related decisions throughout the mate choice process.

In satin bowerbirds, female mate choice occurs in three stages: 'visits', 'pre-nest-building courtships' (pre-NB) and 'post-nest-building courtships' (post-NB)^{11–13}. Before pre-NB courtships, most females (year 1999, 80%; year 2000, 77%) engage in a stage of visits to males' bowers while the bower owners are absent—visits allowed us to evaluate the effect of blue bower decorations on female mate choice decisions independent of a male's intense courtship display. After visits, each female engages in pre-NB courtships at the bowers of several males. During courtship, the female stands inside the bower while the male displays intensely on the platform in front of the bower⁹. After pre-NB courtships, each female spends approximately 1 week building her nest before returning for several

post-NB courtships with a subset of the males previously sampled¹¹. From this subset, each female typically chooses a single male as a mate. After each visit to a bower, a female has two options: reject the male, or return for pre-NB courtship. After each courtship a female has three options: reject the male, return for another courtship, or copulate. Detailed information on individual female's mate-searching patterns allowed us to evaluate which of these options was chosen at each stage of mate choice and whether female age affected the use of blue bower decorations and male display intensity in these decisions.

In the first stage of the mate choice process, females visit males' bowers while the bower-owning males are absent. After each visit, a female must decide whether to reject the bower-owning male from her pool of potential mates or return to him for pre-NB courtship. By evaluating the mean proportion of females that returned to 'experimental' or 'control' males (Fig. 1) for pre-NB courtships after visits, we tested the hypothesis that females use blue decorations in the first stage of mate choice. In both 1999 and 2000, all age classes of females preferentially returned for pre-NB courtships with experimental males over control males (Fig. 2a, b; see also Supplementary Tables S1 and S2). These results show that females of all age classes assess blue bower decorations during visits and use the information in decisions related to returning for pre-NB courtships.

In the second stage of the mate choice process, females engage in pre-NB courtships with multiple males. During these courtships, the bower-owning male is present and all elements of male display, including courtship display intensity, are available for assessment. We tested the hypothesis that blue decorations affect females' decisions to return for post-NB courtships. After pre-NB courtships, we found age-related differences in the proportion of females returning to experimental and control males. First- and second-year females returned preferentially to experimental males for post-NB courtships in both years, whereas there was no effect on the post-NB return rates of three-plus females in either year (Fig. 3a, b; see also Supplementary Tables S1 and S2). In both years, larger proportions of first- and second-year females compared with three-plus females returned to experimental males for post-NB courtships (Fig. 3a, b).

After building their nests, females initiate the third stage of the mate choice process by engaging in post-NB courtships with a subset of the males sampled in pre-NB courtships^{11,12}. From this subset, a female typically chooses a single male as a mate, and we tested the hypothesis that females prefer males with augmented blue decorations as mates. In 1999, second-year and three-plus females did not prefer experimental males as mates, whereas first-year



Figure 1 Photograph of a bower with experimentally augmented blue decorations in 2000. Experimental males placed 20 blue tiles and 50 blue plastic strands on their bower platforms; bower decorations of control males remained unaugmented.

females tended to prefer experimental males as mates (Fig. 4a; see also Supplementary Table S1). In 2000, we increased the number of males in the experiment and increased the number of decorations used to augment bowers (Fig. 1). Again, the mate choices of three-plus females were not affected by the augmentation (Fig. 4b); however, first- and second-year females showed significant mating preferences for experimental males (Fig. 4b; see also Supplementary Table S2). Among age classes in 1999, there were no differences in the proportion of females that chose experimental males as mates, whereas in 2000, larger proportions of first- and second-year females compared with three-plus females chose experimental males as mates (Supplementary Table S7).

Our study shows that young females place an emphasis on blue decorations in decisions made throughout the mate choice process, whereas older females use blue decorations in decisions only when a male's behavioural displays are unavailable for assessment. A previous study suggested that young females are more threatened and less tolerant of intense male displays than older females¹⁰, and therefore may be averse to using intense male displays in mate choice. In 2000, we tested the hypothesis that the mate choice

decisions of older females are affected by male display intensity, whereas the decisions of younger females are not. There was no difference in mean display intensity between experimental and control males (experimental males = $3,130.25 \pm 189.21$; control males = $2,949.33 \pm 210.04$; $t_{26} = 0.64$, $P = 0.53$), therefore, we considered all males as a single group for regression analyses. As predicted, display intensity did not explain a significant amount of the variation in the proportion of first-year females ($r^2 = 0.07$, $F_{1,20} = 1.67$, $P = 0.21$) or second-year females ($r^2 = 0.003$, $F_{1,12} = 0.04$, $P = 0.83$) that returned for post-NB courtships. In contrast, male display intensity explained a significant amount of the variation in the proportion of three-plus females that returned for post-NB courtships ($r^2 = 0.38$, $F_{1,26} = 15.67$, $P < 0.01$). Comparisons of regression coefficients among age classes showed that male display intensity had a larger effect on the decisions of three-plus females than younger females ($F_{2,58} = 29.30$, $P < 0.001$). These results are consistent with those from the decoration augmentation experiment: young females place an emphasis on blue decorations throughout the mate choice process, whereas older females use blue decorations only when a

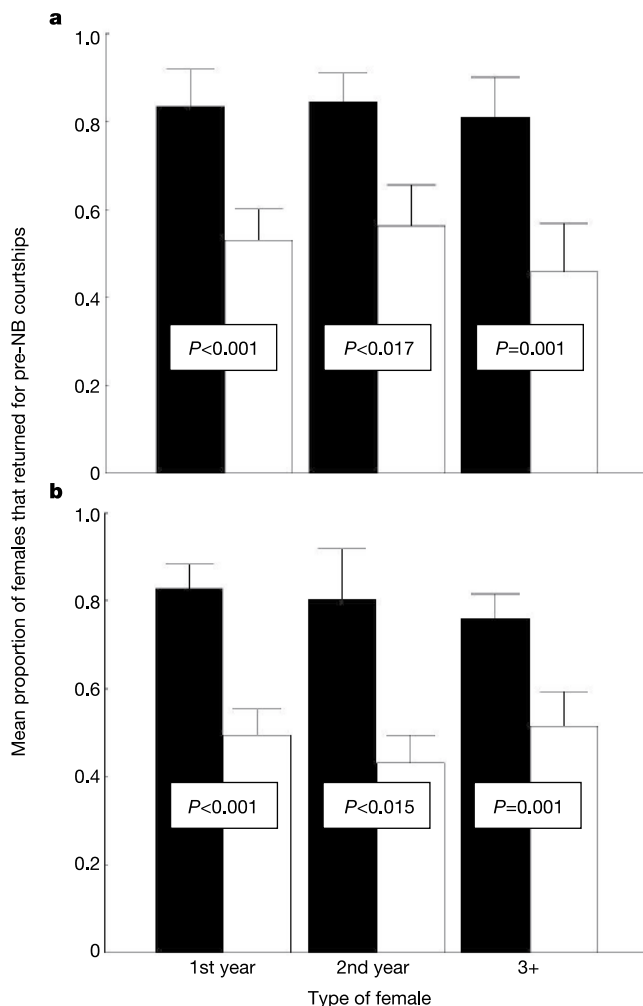


Figure 2 Mean proportion of females that returned to experimental (filled bars) or control (open bars) males for pre-NB courtships in 1999 (a) and 2000 (b). In both years, after visits all females preferentially returned to experimental males for pre-NB courtships. Bars and error bars represent mean + standard error. P values for within-age class comparisons are inset. Among age classes, there were no differences in the mean proportion of females that returned to experimental males for pre-NB courtships. For complete t values, degrees of freedom and P values, see Supplementary Table S4.

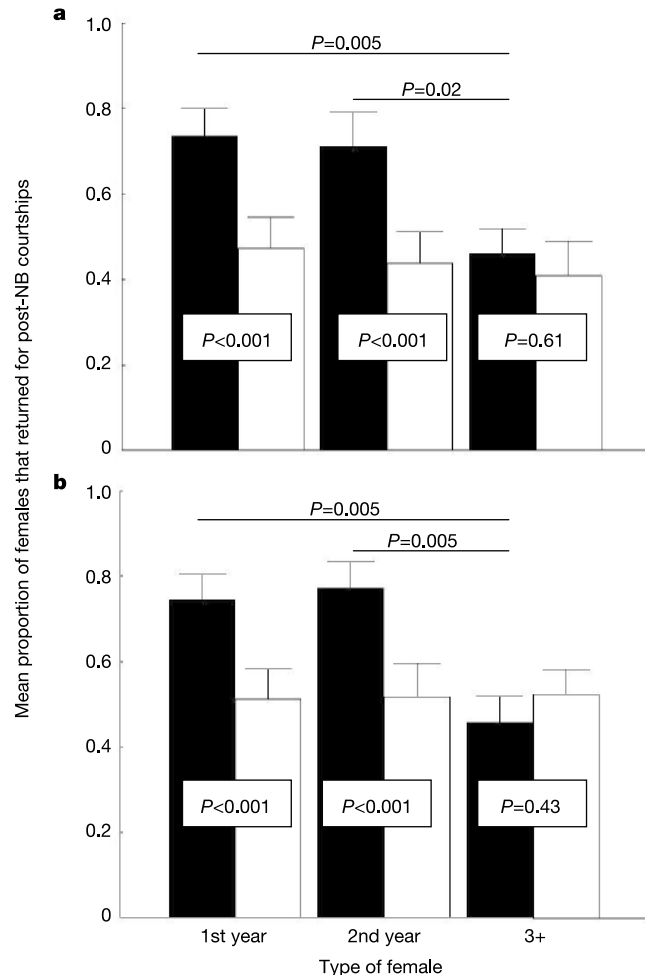


Figure 3 Mean proportion of females that returned to experimental (filled bars) or control (open bars) males for post-NB courtships in 1999 (a) and 2000 (b). In both years, after pre-NB courtships first- and second-year females preferentially returned to experimental males for post-NB courtships; three-plus females did not. Bars and error bars represent mean + standard error. P values for within-age class comparisons are inset. P values at the top of each panel indicate significant differences among age classes in the mean proportion of females that returned to experimental males for post-NB courtship. For complete t values, degrees of freedom and P values, see Supplementary Table S5.

male's behavioural display elements, such as display intensity, are unavailable for assessment. Age-related differences in females' tolerance of intense male displays¹⁰ may explain the age-related reliance on different display elements in mate choice. An alternative explanation is that the differences in old and young females' use of blue decorations could result because of age-related differences in female fidelity^{11,12} but we found no support for this hypothesis (Supplementary Table S3).

We show that mate choice in satin bowerbirds is a complex process made up of multiple stages, where females make sequential stage-specific decisions based on the assessment of male display traits, and there are age-specific differences among females in the male traits used in assessment. The differences in females' use of blue decorations versus display intensity may be related to the threatening nature of male displays. Display intensity may be a better indicator of male quality than are blue decorations^{1,9,10}, but because young females are more easily threatened by intense displays than are older females¹¹, young females may be unable to assess display intensity during mate choice. Similar age- and stage-specific female preferences may occur in the numerous other species where males have multifaceted displays, where females sample multiple

males before choosing mates, and where females choose mates based on the assessment of multiple display traits^{1,14–17}. The variation in female preferences may not have been detected previously because detection requires experimental manipulation of male traits combined with intensive monitoring of individual females throughout the mate choice process. We suggest that as more intensive studies of mate choice are done in which intrinsic female factors (that is age/experience) and the stages of mate choice are considered, variable female preferences will be discovered and may help explain the widespread occurrence of multifaceted male displays. □

Methods

Marking and monitoring

This work was conducted in 1999 and 2000 at Wallaby Creek, New South Wales, Australia. Before the mating season, birds were captured using traps and mist nets. Each individual was fitted with three plastic leg bands arranged in a unique colour combination. In our study population, all adult males and most (1999, 87%; 2000, 91%) of the females were uniquely marked for identification. From 1 November to 20 December, automatic Hi-8 video cameras recorded behaviours at 24 (1999) and 28 (2000) adjacent bowers. This monitoring provided a complete record of all courtships and copulations. Individuals were only classified as female if they were observed copulating with a male on videotape. Detailed mate-sampling patterns and mate choice for 52 (1999) and 64 (2000) females were reconstructed from video footage.

Decoration augmentation experiment

Using information from 1998 (refs 12,13), bower-owning males were paired to maximize similarity in mating success (number of different mates). In 1999, the bower of one male in each dyad was randomly selected for augmentation with blue bower decorations (experimental males), while the other male's decorations remained unaugmented (control males). In 1999, 20 blue plastic tiles (2.54 cm²) were placed in two arcs of 10 tiles on the bower platform of each experimental male. In 2000, we used the same males in the experimental and control groups in addition to two new experimental males, paired with two new control males. In this second year, 20 blue tiles and 50 strands of blue plastic (25 × ~0.20 cm) (Fig. S1) were placed in caches 1 m from the bower platform. Within 2 h, each experimental male placed all of the tiles and plastic strands on his bower platform (Fig. 1). Blue plastic strands are frequently used as bower decorations in the Wallaby Creek population, and the number of strands used in the augmentation was within the range of plastic strands displayed by bower-owning males before the augmentation (mean number ± standard error of plastic strands on bowers before the experiment, 45 ± 7.73) (S.W.C. and G.B., manuscript in preparation). To prevent decoration stealing¹⁸, after experimental males placed their tiles each tile was glued to the head of a long screw then secured into the ground. Plastic strands were frequently woven into the bower platform by bower owners, probably reducing the frequency of stealing (S.W.C. and G.B., personal observations). Twice a day at each bower, the number of tiles and plastic strands were counted and replaced to original levels if necessary. Tiles and plastic strands found on the bowers of control males as a result of decoration stealing¹⁸ were removed. All other bower decorations⁸ remained unmanipulated.

Female age classes

Females were grouped based on their years of mating experience. Three-plus females had at least 2 yr of previous mating experience. Second-year females had 1 yr of previous mating experience. First-year females were not previously observed mating. All first-year females in 1999 (*n* = 12) and 2000 (*n* = 16) were captured and marked for the first time in those years. Further suggesting that these were first-year females, we found that they weighed significantly less than three-plus females (*t*₆₅ = 2.11, *P* < 0.01). All females used in our analyses were courted by at least one experimental male and one control male.

Male display intensity

Intense male displays involve loud buzzing vocalizations, piloerection and vigorous running across the bower platform with wings extended, all performed in close proximity to the female⁹. For each male, we calculated display intensity in the first pre-NB courtship with each courted female in 2000. We calculated display intensity based on four variables: (1) the distance run by the male during his display; (2) the degree of male piloerection, appearing larger; (3) the location of the male on the bower platform; and (4) the degree of the male's wing extension, appearing larger^{9,10}. For each male we calculated mean display intensity, and used these means in regression analyses to evaluate the relative contribution of display intensity to the proportion of females returning for post-NB courtships.

Statistical analyses

To account for the unequal numbers of females sampling each male in visits and pre-NB courtships (Supplementary Tables S1 and S2), we treated each male as a replicate and weighted the proportion of returning females by the inverse of the variances in the proportion of females observed in visits and pre-NB courtships.

Received 21 November 2003; accepted 11 February 2004; doi:10.1038/nature02419.

- Andersson, M. *Sexual Selection* (Princeton Univ. Press, Princeton, 1994).
- Borgia, G. Complex male display and female choice in the spotted bowerbird: specialized functions for different bower decorations. *Anim. Behav.* **49**, 1291–1301 (1995).

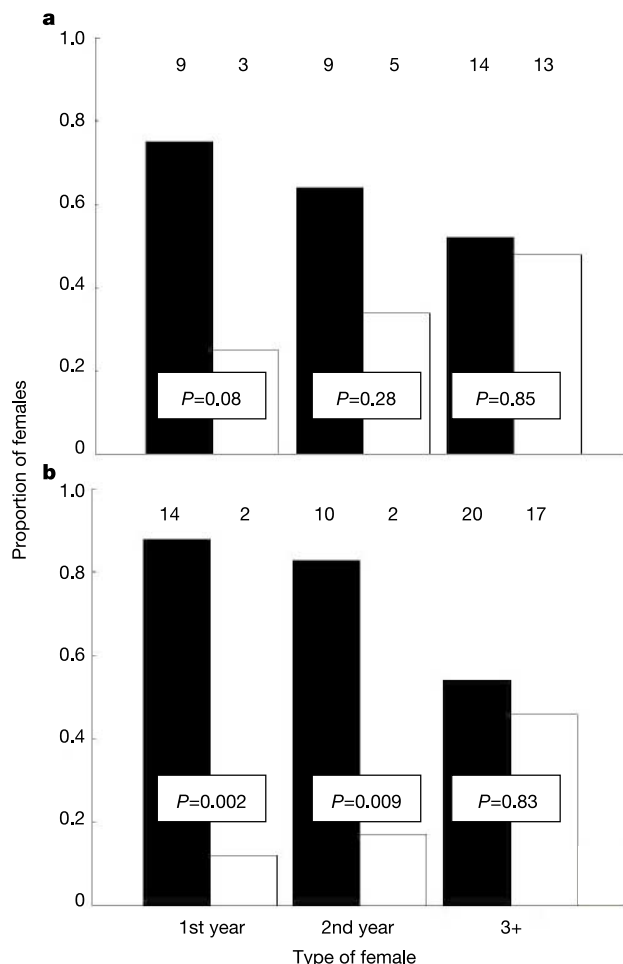


Figure 4 Proportion of females that mated with experimental males (filled bars) or control males (open bars) in 1999 (a) and 2000 (b). In 1999, first-year females tended to prefer experimental males as mates; second-year and three-plus females did not. In 2000, first- and second-year females preferred experimental males; three-plus females did not. *P*-values for within-age class comparisons are inset. For χ^2 values and degrees of freedom, see Supplementary Table S6. For comparisons among age classes in the proportion of females that chose experimental males as mates, see Supplementary Table S7. Sample sizes are indicated above the bars.

3. Møller, A. P. & Pomiankowski, A. Why have birds got multiple sexual ornaments. *Behav. Ecol. Sociobiol.* **32**, 167–176 (1993).
4. Johnstone, R. A. Multiple displays in animal communication: 'backup signals' and 'multiple messages'. *Phil. Trans. R. Soc. Lond. B* **351**, 329–338 (1996).
5. Brooks, R. & Endler, J. A. Female guppies agree to differ: phenotypic and genetic variation in mate-choice behavior and the consequences for sexual selection. *Evolution* **55**, 1644–1655 (2001).
6. Kodric-Brown, A. & Nicoletto, P. F. Age and experience affect female choice in the guppy (*Poecilia reticulata*). *Am. Nat.* **157**, 316–323 (2001).
7. Moore, P. J. & Moore, A. J. Reproductive aging and mating: The ticking biological clock in female cockroaches. *Proc. Natl Acad. Sci. USA* **98**, 9171–9176 (2001).
8. Borgia, G. Bower quality, number of decorations and male mating success of male satin bowerbirds (*Ptilonorhynchus violaceus*): an experimental analysis. *Anim. Behav.* **33**, 266–271 (1985).
9. Patricelli, G. L., Uy, J. A. C., Walsh, G. & Borgia, G. Sexual selection: male displays adjusted to female's response. *Nature* **415**, 279–280 (2002).
10. Patricelli, G. L., Uy, J. A. C. & Borgia, G. Female signals enhance the efficiency of mate assessment in satin bowerbirds (*Ptilonorhynchus violaceus*). *Behav. Ecol.* **15**, 297–304 (2004).
11. Uy, J. A. C., Patricelli, G. L. & Borgia, G. Complex mate searching in the satin bowerbird (*Ptilonorhynchus violaceus*). *Am. Nat.* **158**, 530–542 (2001).
12. Uy, J. A. C., Patricelli, G. L. & Borgia, G. Dynamic mate searching tactic allows female satin bowerbirds (*Ptilonorhynchus violaceus*) to reduce searching. *Proc. R. Soc. Lond. B* **267**, 251–256 (2000).
13. Uy, J. A. C., Patricelli, G. L. & Borgia, G. Loss of preferred mate forces female satin bowerbirds (*Ptilonorhynchus violaceus*) to increase mate searching. *Proc. R. Soc. Lond. B* **268**, 633–638 (2001).
14. Widemo, F. & Sæther, S. A. Beauty is in the eye of the beholder: causes and consequences of variation in mating preferences. *Trends Ecol. Evol.* **14**, 26–31 (1999).
15. Jennions, M. D. & Petrie, M. Variation in mate choice and mating preferences: a review of causes and consequences. *Biol. Rev.* **72**, 283–327 (1997).
16. Gibson, R. M. & Langen, T. A. How do animals choose their mates. *Trends Ecol. Evol.* **11**, 468–470 (1996).
17. Höglund, J. & Alatalo, R. *Leks* (Princeton Univ. Press, Princeton, 1995).
18. Borgia, G. & Gore, M. A. Sexual competition by feather stealing in the satin bowerbird (*Ptilonorhynchus violaceus*). *Anim. Behav.* **34**, 727–738 (1986).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank G. S. Wilkinson, K. L. Shaw, J. A. C. Uy and T. C. Mendelson for comments on the manuscript, and M. C. Christman for statistical advice. We also thank the Australian Bird and Bat Banding Scheme, New South Wales National Parks and Wildlife Service, the landowners at Wallaby Creek, and the 1998–2000 field assistants. This research was funded by the National Science Foundation, Animal Behavior Program (USA) to G.B., and by fellowships from the Behavior, Ecology, Evolution and Systematics Program of the University of Maryland, and the N.S.F. Research Training Grant to S.W.C. and G.L.P.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.W.C. (Coleman@umd.edu).

Variation in behaviour promotes cooperation in the Prisoner's Dilemma game

John M. McNamara¹, Zoltan Barta¹ & Alasdair I. Houston²

¹Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, UK

²School of Biological Sciences, University of Bristol, Woodland Road, Bristol BS8 1UG, UK

The Prisoner's Dilemma game^{1–4} is widely used to investigate how cooperation between unrelated individuals can evolve by natural selection. In this game, each player can either 'cooperate' (invest in a common good) or 'defect' (exploit the other's investment). If the opponent cooperates, you get R if you cooperate and T if you defect. If the opponent defects, you get S if you cooperate and P if you defect. Here $T > R > 0$ and $P > S$, so that 'defect' is the best response to any action by the opponent. Thus in a single play of the game, each player should defect. In our game, a fixed maximum number of rounds of the Prisoner's Dilemma game is played against the same opponent. A standard argument based on working backwards from the last round^{1,5} shows that

defection on all rounds is the only stable outcome. In contrast, we show that if extrinsic factors maintain variation in behaviour, high levels of co-operation are stable. Our results highlight the importance of extrinsic variability in determining the outcome of evolutionary games.

In our version of the Prisoner's Dilemma game, the fixed maximum number of possible rounds, N , is known to both players. On each round before the last one, if either player defects then the game ends; only if both cooperate do they proceed to the next round. (This approach reflects the ability of mobile organisms to terminate an unfavourable interaction by leaving^{6,7}.) After N rounds, the interaction ends whatever decisions are made. The total payoff is the sum of the payoff from each round.

We assume that $T > (P + R)$. The standard arguments based on working backwards then lead to the conclusion that defection on all rounds is the only evolutionarily stable outcome in our game (see Supplementary Information). At evolutionary stability all population members behave in this way, so that there is no variation in the population, and unlike some iterated Prisoner's Dilemma games, variation cannot be maintained by frequency dependence. We believe, however, that in real populations there are always factors other than frequency dependence—such as mutation, immigration, recombination and epistasis—that maintain genetic variation. We show that maintaining variation fundamentally changes the nature of the game. The intuition behind this is as follows. In a population at the defect evolutionarily stable strategy, a player should defect on the first round. But if variability is maintained, and hence there is a chance that an opponent will cooperate, then there is the potential for a substantial gain, and it may be worth cooperating initially in the hope that the opponent is cooperative (compare ref. 1). Whether for this or for other reasons, humans do not defect as much as expected in the Prisoner's Dilemma and related games^{5,8}.

In our model, a strategy specifies the number of rounds n to cooperate before defecting. (The game may not last for $n + 1$ rounds because the opponent may terminate the game by defecting beforehand.) We consider the evolution of an infinite population with discrete generations. The number of offspring left by an individual in the next generation is the individual's total payoff plus a positive constant, K , which represents the contributions to fitness that come from outside the game. If a parent adopts strategy n then each offspring is also n with probability $1 - 2\epsilon$. Genetic variability is maintained by mutation; with probability ϵ offspring are $n - 1$ and with probability ϵ they are $n + 1$. We use the standard

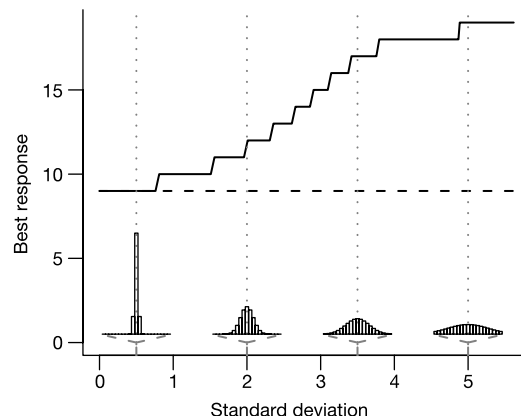


Figure 1 The best response (continuous line) as a function of the variation in the degree of cooperation in the population. In all cases, the mean number of rounds of cooperation before defection is $E(n) = 10$. The distribution of n about this mean is pseudo-normal with the standard deviation indicated. Four actual distributions are illustrated. The dashed line gives the best response when each round is chosen against a new opponent, randomly selected from the population. In these calculations, $N = 20$.

payoffs for the Prisoner's Dilemma game as our baseline, that is, $T = 5$, $R = 3$, $P = 1$, $S = 0$.

If you know that the other player will cooperate for n rounds and defect on the $(n + 1)$ th round, then the best thing to do is to preempt this defection by cooperating for $n - 1$ rounds and defecting on the n th round (see Supplementary Information). It follows that if all members of the population adopt the strategy n , then the best response (that is, the best value of n) for a mutant is $n - 1$. The situation changes if the population consists of a mixture of individuals each with different strategies. If the population mean is $E(n)$, then the best response can be greater than $E(n)$ when there is variation in the population (see Supplementary Information). This is because some population members cooperate for more than $E(n)$ rounds, and the potential benefit of interacting with such an opponent outweighs the cost incurred if the opponent is the first to defect. Figure 1 shows the best response for a mutant in a population for which $E(n) = 10$. As the variation about the mean increases, so does the best response. This effect is not restricted to the game that we consider; it can occur in any game in which the benefit of interacting with a cooperative opponent is significantly greater than the cost of being exploited by an uncooperative opponent. Box 1 gives an analytic example based on a continuous trait.

Figures 2 and 3 illustrate the evolutionarily stable outcomes when a population evolves from an initial population that consists entirely of animals that defect on the first round. (Computations show that the same stable outcome is reached from any initial population.) Fig. 2 shows that the outcome depends on the

mutation rate ε . If ε is below a critical value, then the population is essentially non-cooperative, with a small amount of cooperation maintained by mutation. In contrast, if ε is above a critical value, then the population is essentially cooperative, but with some variation. In this case the unique best response to the population is more cooperative than the mean level of cooperation. Cooperation does not, however, increase further because of biased mutation at the upper limit of possible cooperation. As Fig. 3 shows, the sharp transition from a non-cooperative solution to a cooperative solution as ε increases is a general feature of our model.

Figure 3a shows the effect of the parameter K . The bigger the background fitness contribution K , the less important is the payoff from the game in determining an individual's lifetime fitness, and hence the more genetic variation there is likely to be in a population for a given value of ε . Thus when K is large, the transition to cooperation occurs at a lower value of ε . This means that we predict that cooperation becomes less likely to evolve as the game payoff constitutes a greater proportion of the lifetime reproductive success of the organism. Figure 3b shows the effect of the maximum possible number of rounds, N . The transition to cooperation is almost independent of N , provided that $N > 2$. Sensitivity analysis reveals that our conclusions are robust to changes in the values of the payoffs R , P , S and T .

We have illustrated our argument using mutation as a source of genetic variation, but similar effects are found if phenotypic variation is produced by errors in decision-making or developmental noise (see Fig. 2b). In natural populations, variation is likely to be maintained by multiple factors.

Box 1

Mutation rate determines the direction of evolution

We consider a large (infinite) population in which individuals differ in their cooperativeness—a continuous trait. Population members meet at random and engage in a contest. The payoff to an individual with cooperativeness x in a contest with an opponent with cooperativeness y is:

$$w(x, y) = \exp \left\{ \alpha y - \frac{1}{2} (y - 1 - x)^2 \right\}$$

This is maximized when the individual is less cooperative than the opponent by one unit. The other key feature is that simultaneously increasing the cooperativeness of the focal individual and the opponent, while maintaining a constant difference between the two, increases the focal individual's payoff. The rate of increase increases with the parameter α .

The fitness of an individual with cooperativeness x , $W(x)$, is the average of $w(x, y)$ across the cooperativeness y of population members. Assuming that cooperativeness is normally distributed with mean μ and variance σ^2 in the population, it can be shown that:

$$W(x) \propto \exp \left\{ - \frac{[(x + 1) - (\mu + \alpha \sigma^2)]^2}{2(\sigma^2 + 1)} \right\}$$

This is maximized when $x = x^*$, where $x^* = \mu - 1 + \alpha \sigma^2$.

Thus if all population members have cooperativeness $y = \mu$ (that is, $\sigma^2 = 0$), fitness is maximized when $x^* = \mu - 1$. However, if the population variance σ^2 exceeds $1/\alpha$, x^* exceeds the population mean μ .

Evolution is modelled by discrete-generation replicator dynamics, with error. An individual of cooperativeness x leaves $W(x)$ offspring in the next generation. The distribution of offspring cooperativeness is normal, with mean x and variance β^2 . Then, given that cooperativeness is normally distributed in generation t , it is also normally distributed in generation $t + 1$, with successive means and variances related by $\mu_{t+1} = \mu_t + v_t(\alpha \sigma_t^2 - 1)$ and $\sigma_{t+1}^2 = \beta^2 + v_t(\sigma_t^2 + 1)$, where $v_t = \sigma_t^2 / (2\sigma_t^2 + 1)$. Over time, the variance tends to the limiting value $\sigma_\infty^2 = \beta^2 + \beta \sqrt{\beta^2 + 1}$. At this limit:

$$\mu_{t+1} > \mu_t \Leftrightarrow \beta^2 > \frac{1}{\alpha(\alpha + 2)}$$

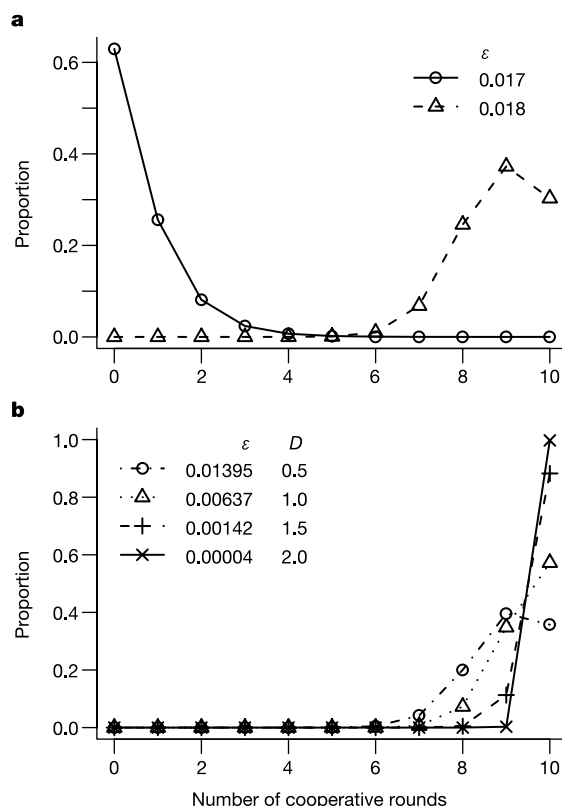


Figure 2 The proportion of each type of individual (as specified by n) in the population at evolutionary stability. **a**, Results from the model presented in the text. Two cases, corresponding to two values of the mutation parameter ε , are shown. **b**, Results when there is also a distribution of phenotypes (with standard deviation D) for a given genotype. Results are presented for four combinations of ε and D . For each value of D , the corresponding value of ε is the minimum at which cooperation occurs. $N = 10$ and $K = 20$ throughout.

Maintaining a small number of defecting individuals in a population playing the iterated Prisoner's Dilemma can influence the evolutionary outcome^{9,10}. In the examples of refs 9 and 10, a population in which individuals cooperate is not evolutionarily stable in a strict sense because several strategies do equally well and their proportions can change by drift. Introducing small numbers of defectors into the population removes this neutrality, preventing the drift towards strategies that could be exploited by defectors. In contrast, our model does not involve neutrality. When a population is at the equilibrium where all individuals defect immediately, any other strategy is strictly worse. Introducing variation changes the fitness landscape so that strategies that initially cooperate are now on the other side of a valley and do better than the strategy of always defecting (Fig. 1). At the stable equilibrium with cooperation (Figs 2 and 3) there is still no neutrality. Instead, there is a unique best response to the population that is more cooperative than the mean level of cooperation.

Our version of the Prisoner's Dilemma game has a similar structure to the centipede game^{5,8}, in that non-cooperation ends the game and cooperating with an opponent that cooperates leads to a high payoff. As in our game, non-cooperation on the first round is the only stable outcome. It is well documented, however, that humans do not adopt this strategy^{8,11}. Models that attempt to account for this behaviour include a certain fraction of completely cooperative individuals in the population^{8,11}. Unlike our model, this fraction is imposed by the modeller rather than emerging from the evolutionary dynamics. It is clear that applying our approach to this game would lead to the evolution of high levels of cooperation even if the mutation rates are very low.

Sources of variation in models of cooperation include stochastic strategies, errors in decision making, mutations and differences in quality^{9,12–23}. All these sources of variation may favour cooperation.

When variability is maintained by mutation, as in our model, not all strategies present at evolutionary stability maximize fitness. This means that the outcome of evolution cannot be found by the standard approach in which the endpoint is a best response to itself.

The strategies that are used to play the Prisoner's Dilemma game are often characterized in terms of personality traits (for example, 'Tit-for-tat' is said²⁴ to be nice, can be provoked but forgives). In our game, players differ in a single aspect of personality: their degree of cooperation. Because a series of rounds is played against the same individual, a player can learn about the opponent's persistence in cooperating. If the opponent has cooperated for n rounds, the probability that it will cooperate on round $n + 1$ is greater than the probability that a randomly chosen population member will cooperate on round $n + 1$. Repeated play against the same opponent is crucial for the evolution of cooperation in our model. If the opponent on each round is chosen anew, then cooperation will not evolve (see Fig. 1).

A possible explanation for the diversity of personality types within social groups^{7,25} is that each type has equal fitness and variation is maintained by frequency dependence (compare alternative male mating strategies^{26,27}, producers and scroungers²⁸). Another possibility is that the success of the group is crucial^{28–30}, and this success depends on the range of types. In contrast, our approach assumes that variability is maintained by various genetic and environmental factors. As a result, a population will always have a range of personalities. Our argument is that an adequate model of the evolution of not just cooperation, but social behaviour in general, should take this into account. Our results show that the outcome may be very different from the outcome when variability is not maintained by extrinsic factors. □

Received 8 October 2003; accepted 18 February 2004; doi:10.1038/nature02432.

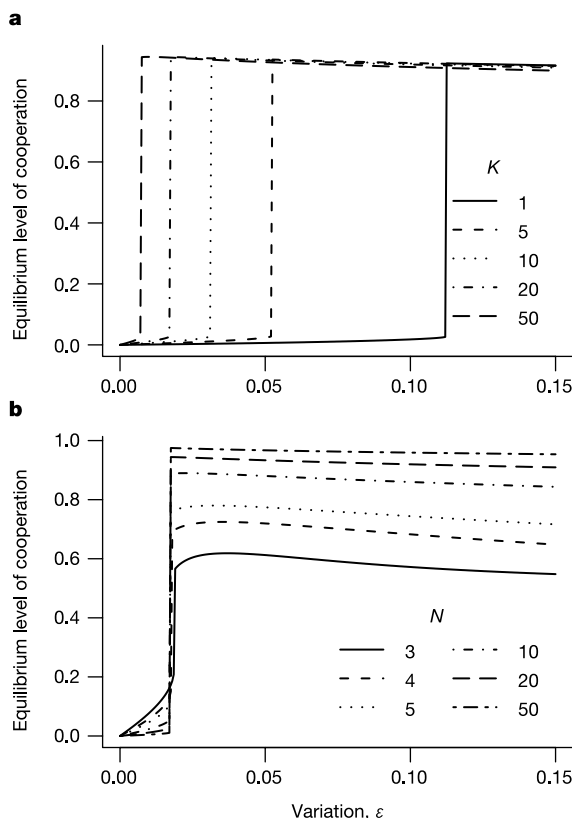


Figure 3 Levels of cooperation in the population, measured as $E(n)/N$, at evolutionary stability, shown as a function of the mutation parameter ε . **a**, Various values of K for the case $N = 20$. **b**, Various values of N for the case $K = 20$.

- Luce, R. D. & Raiffa, H. *Games and Decisions* (Wiley, New York, 1957).
- Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
- Axelrod, R. & Dion, D. The further evolution of cooperation. *Science* **242**, 1385–1390 (1988).
- Brembs, B. Chaos, cheating and cooperation: Potential solutions to the Prisoner's Dilemma. *Oikos* **76**, 14–24 (1996).
- Colman, A. M. Cooperation, psychological game theory, and the limitations of rationality in social interaction. *Behav. Brain Sci.* **26**, 139–198 (2003).
- Enquist, M. & Leimar, O. The evolution of cooperation in mobile organisms. *Anim. Behav.* **45**, 747–757 (1993).
- Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425**, 785–791 (2003).
- Camerer, C. *Behavioral Game Theory* (Princeton Univ. Press, Princeton, 2003).
- Crowley, P. H. & Sargent, R. C. Whence tit-for-tat? *Evol. Ecol.* **10**, 499–516 (1996).
- Sherratt, T. N. & Roberts, G. The importance of phenotypic defectors in stabilizing reciprocal altruism. *Behav. Ecol.* **12**, 313–317 (2001).
- McKelvey, R. D. & Palfrey, T. R. An experimental study of the centipede game. *Econometrica* **60**, 803–836 (1992).
- Boyd, R. Mistakes allow evolutionary stability in the repeated Prisoner's Dilemma game. *J. Theor. Biol.* **136**, 47–56 (1989).
- Fudenberg, D. & Maskin, E. Evolution and cooperation in noisy repeated games. *Am. Econ. Rev.* **80**, 274–279 (1990).
- Nowak, M. & Sigmund, K. The evolution of stochastic strategies in the Prisoner's Dilemma. *Acta Appl. Math.* **20**, 247–265 (1990).
- Bendor, J., Kramer, R. M. & Stout, S. When in doubt. Cooperation in a noisy prisoner's dilemma. *J. Conflict Resolut.* **35**, 691–719 (1991).
- Nowak, M. & Sigmund, K. Tit for tat in heterogeneous populations. *Nature* **355**, 250–252 (1992).
- Nowak, M. & Sigmund, K. Chaos and the evolution of cooperation. *Proc. Natl Acad. Sci. USA* **90**, 5091–5094 (1993).
- Nowak, M. & Sigmund, K. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature* **364**, 56–58 (1993).
- Nowak, M. A., Sigmund, K. & Eshed, E. Automata, repeated games and noise. *J. Math. Biol.* **33**, 703–722 (1995).
- Wu, J. Z. & Axelrod, R. How to cope with noise in the iterated prisoner's dilemma. *J. Conflict Resolut.* **39**, 183–189 (1995).
- Wahl, L. M. & Nowak, M. A. The continuous prisoner's dilemma: II. Linear reactive strategies with noise. *J. Theor. Biol.* **200**, 323–338 (1999).
- Lorberbaum, J. P., Bohning, D. E., Shastri, A. & Sine, L. E. Are there really no evolutionarily stable strategies in the iterated prisoner's dilemma? *J. Theor. Biol.* **214**, 155–169 (2002).
- Fishman, M. A., Lotem, A. & Stone, L. Heterogeneity stabilizes reciprocal altruism interactions. *J. Theor. Biol.* **209**, 87–95 (2001).
- Axelrod, R. More effective choice in the Prisoner's Dilemma. *J. Conflict Resolut.* **24**, 379–403 (1980).
- Heinsohn, R. & Packer, C. Complex cooperative strategies in group-territorial African lions. *Science* **269**, 1260–1262 (1995).
- Brockmann, H. J. The evolution of alternative strategies and tactics. *Adv. Stud. Behav.* **30**, 1–51 (2001).

27. Henson, S. A. & Warner, R. R. Male and female alternative reproductive behaviors in fishes: a new approach using intersexual dynamics. *Annu. Rev. Ecol. Syst.* **28**, 571–592 (1997).
28. Giraldeau, L. A. & Caraco, T. *Social Foraging Theory* (Princeton Univ. Press, Princeton, 2000).
29. Sober, E. & Wilson, D. S. *Unto Others* (Harvard Univ. Press, Cambridge, Massachusetts, 1998).
30. Boyd, R. & Richerson, P. J. Group selection among alternative evolutionarily stable strategies. *J. Theor. Biol.* **145**, 331–342 (1990).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. Dornhaus, M. Enquist, E. Fehr and L.-A. Giraldeau for comments on a previous version of this Letter.

Authors' contributions J.M.M. formulated the main ideas as a result of conversations with A.I.H.; J.M.M. also formulated the model, and was responsible for the material in Box 1; Z.B. carried out the computations, and prepared the figures; A.I.H. surveyed the literature, and had the main responsibility for writing the Letter.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.M.M. (john.mcnamara@bristol.ac.uk).

Neural activity predicts individual differences in visual working memory capacity

Edward K. Vogel & Maro G. Machizawa

Department of Psychology, University of Oregon, Eugene, Oregon 97403-1227, USA

Contrary to our rich phenomenological visual experience, our visual short-term memory system can maintain representations of only three to four objects at any given moment^{1,2}. For over a century, the capacity of visual memory has been shown to vary substantially across individuals, ranging from 1.5 to about 5 objects^{3–7}. Although numerous studies have recently begun to characterize the neural substrates of visual memory processes^{8–12}, a neurophysiological index of storage capacity limitations has not yet been established. Here, we provide electrophysiological evidence for lateralized activity in humans that reflects the encoding and maintenance of items in visual memory. The amplitude of this activity is strongly modulated by the number of objects being held in the memory at the time, but approaches a limit asymptotically for arrays that meet or exceed storage capacity. Indeed, the precise limit is determined by each individual's memory capacity, such that the activity from low-capacity individuals reaches this plateau much sooner than that from high-capacity individuals. Consequently, this measure provides a strong neurophysiological predictor of an individual's capacity, allowing the demonstration of a direct relationship between neural activity and memory capacity.

To measure the neural correlates of visual memory capacity, we recorded event-related potentials (ERPs) from normal young adults while they performed a visual memory task. On each trial they were presented with a brief bilateral array of coloured squares and were asked to remember the items in only one hemifield, which was indicated with an arrow (Fig. 1a). Memory was tested one second later with the presentation of a test array that was either identical to the memory array or differed by one colour. Subjects pressed one of two buttons to indicate whether the two arrays were identical or different. We have used variations of this paradigm previously and have found that observers are accurate for array sizes of up to three

to four items, and that performance is not significantly influenced by perceptual or verbal processes^{1,3}.

In the first experiment, we recorded ERPs to the onset of a four-item memory array so that we could observe the sustained electrophysiological response during the memory retention interval. A few previous ERP studies have observed a sustained response during working memory tasks for foveally presented stimuli, but did not examine lateralized effects^{13,14}. In contrast, we took advantage of the primarily contralateral organization of the visual system by presenting lateralized stimuli in each hemifield so that we could measure the spatially specific hemispheric responses to memory arrays that were either contralateral or ipsilateral with respect to electrode position^{15,16}. Approximately 200 ms after the onset of the memory array, we found a large negative-going voltage over the hemisphere that was contralateral to the memorized hemifield, and this response persisted throughout the duration of the memory retention interval (Fig. 1b). This response was focused primarily over the posterior parietal and lateral occipital electrode sites and strongly resembled delay activity recorded from individual neurons in monkey visual cortex^{12,17}.

Numerous processes contribute to visual memory performance, and we sought to determine which aspects of processing are reflected by the contralateral delay activity. Although this effect seems to reflect the maintenance of object representations from the memory array, it is necessary to rule out the possibility that it reflects executive processes¹⁸ involved in performing the task, or even more general processes such as increased effort or arousal^{19–21}. In the second experiment, we tested this by varying the number of items in the memory array to establish whether the amplitude is sensitive to the number of representations that are being held in visual memory. Memory arrays in this experiment varied from one to four items in each hemifield (average capacity in this task is normally around three items^{3,7}). To compare directly the magnitude of activity across array sizes, we constructed 'difference waves' in which the ipsilateral activity was subtracted from the contralateral activity for each array size, which removes the contribution of any nonspecific, bilateral ERP activity.

As shown in Fig. 2a, the amplitude was highly sensitive to the number of items in the memory array. Indeed, increasing an array

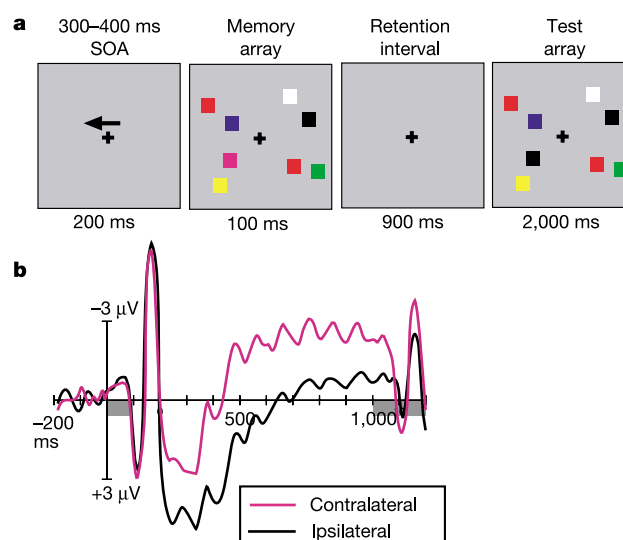


Figure 1 Stimuli and results from experiment one. **a**, Example of a visual memory trial for the left hemifield. SOA, stimulus onset asynchrony. **b**, Grand averaged ERP waveforms time-locked to the memory array averaged across the lateral occipital and posterior parietal electrode sites in experiment one. The two grey rectangles reflect the time periods for the memory and test arrays, respectively. Note that, by convention, negative voltage is plotted upwards.

from one to two squares or from two to three squares resulted in a substantial increase in amplitude. Moreover, because memory performance for near-capacity arrays can fluctuate over time, leading to occasional incorrect responses, we compared the amplitude of the delay activity for correct and incorrect trials. The amplitude for incorrect trials was considerably smaller than that for correct trials ($P < 0.01$), further suggesting that the delay activity specifically reflects the maintenance of successful representations in visual memory. Nevertheless, it is possible that the extent of executive processes also increases with additional memory items. Moreover, there are small but reliable differences in accuracy across array sizes, which leaves open the possibility that increases in arousal or effort for larger arrays may have produced the increase in amplitude.

The amplitude of the contralateral delay activity may have increased as the result of increasing the number of representations, more executive processing, or higher difficulty; however, these alternatives make different predictions for array sizes that exceed visual memory capacity. For example, when comparing a trial containing four memory items to a trial containing eight, the number of active memory representations should be approximately identical, because both trials exceed a typical individual's memory capacity. That is, the subject can maintain only three to four items whether the attended side of the array contains four or eight items. In contrast, the difficulty and extent of executive processing increases substantially for eight-item arrays compared with four-item arrays²². Indeed, this has been a significant limitation of previous neurophysiological studies that have reported memory load effects, because the amount of activity continues to increase for loads that exceed capacity, indicating that these measures are not directly measuring memory capacity^{10,21,23}. Therefore, in the third and fourth experiments, we compared the delay activity for supra-capacity arrays with memory arrays at or near capacity. If it reflects the active representations held in visual memory, we would expect no difference in amplitude between supra-capacity arrays and

capacity arrays. However, if it reflects executive processes or the amount of general effort, we would expect that amplitude should continue to increase for supra-capacity arrays.

The results of the third experiment show that although there was a significant increase in amplitude from arrays of two items per side to arrays of four items per side, there was no increase from four items to six items (Fig. 2b). That is, the amplitude reached a limit with arrays of approximately four items per side. We tested this further in the fourth experiment by following the same experimental design but with larger array sizes (Fig. 2c). Again we found a significant amplitude increase from two to four items per side, but no increase from four items to either eight or ten items per side. These results strongly support the hypothesis that the delay activity reflects the specific maintenance of representations in visual memory because its amplitude is sensitive to the number of successful representations that are active in memory at the time. In addition, the absence of continued amplitude increase beyond capacity also minimizes the possibility that the sub-capacity amplitude effects in

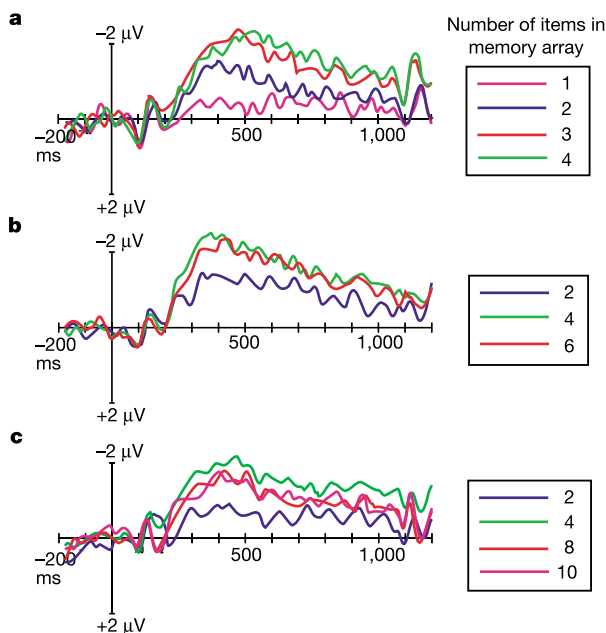


Figure 2 ERP difference waves at lateral occipital and posterior parietal electrode sites for experiments two, three and four, respectively. **a**, Pairwise comparisons yielded significant differences in amplitude between array sizes of one, two and three ($P < 0.001$), but no difference between three and four items ($P > 0.20$) in experiment two. **b, c**, No significant differences in amplitude were observed between arrays of four, six, eight or ten items ($P > 0.25$ in all cases) in experiments three and four.

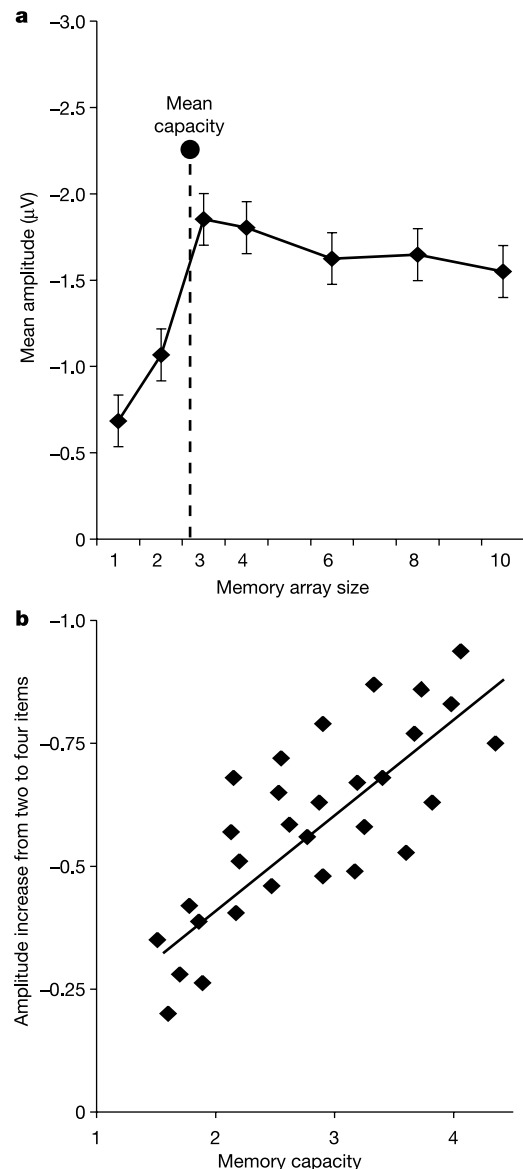


Figure 3 Mean amplitude and visual memory capacity. **a**, Mean amplitude and visual memory capacity across experiments two, three and four. Error bars reflect 95% confidence intervals. **b**, The correlation between an individual subject's memory capacity and the increase in amplitude of delay activity between two- and four-item arrays.

the second experiment were because of increases in the size of the 'attentional spotlight'²⁴, because supra-capacity arrays require a larger spotlight than at- or below-capacity arrays, but show no increases in amplitude.

The supra-capacity array sizes in these experiments provided substantial increases in both the extent of executive processes and the difficulty in performing the task. For example, there was a 32% reduction in accuracy between arrays of four and ten items, yet there was no increase in the amplitude of the contralateral delay activity. Furthermore, we also observed a more centrally distributed bilateral wave during the task that was modulated by the number of items in the memory array. However, in sharp contrast to the contralateral activity, the amplitude of this bilateral wave continued to increase significantly for arrays that exceeded memory capacity, suggesting that it is sensitive to the amount of general effort involved in performing the task²¹.

These results suggest that the contralateral delay activity indexes the currently active representations maintained in visual memory; that is, increasing in magnitude as the number of items increases, but reaching a limit once visual memory capacity is exhausted. To demonstrate this effect further, we quantified the mean amplitudes for each array size for experiments two to four. As shown in Fig. 3a, amplitude increased monotonically from one to three items, but this increase levelled off at three items. We also computed visual memory capacity estimates for each subject, using a standard formula^{7,25}. The mean capacity of the group was 2.8 items, which is approximately when the memory delay activity reaches a limit. This further supports the proposal that the specific limitation in visual memory capacity determines when this delay activity reaches a limit.

To gauge this relationship more finely, we examined the variability across individuals for each measure. That is, we assessed whether a given individual's memory capacity specifically dictates when his or her delay activity reaches a limit. If so, one would expect low-capacity subjects to reach the limit for smaller array sizes than would high-capacity subjects. Unfortunately, it is difficult to determine precisely the limit with a categorical data set such as array size (for example, there is no array size of 2.6). Instead, we computed the amplitude increase between two items and four items per side for each subject across all experiments, the logic being that the amount of amplitude increase between these two array sizes should be specifically determined by memory capacity. For example, when a subject with a low capacity of 1.8 items is shown a two-item array, capacity should be completely consumed with the amplitude reaching a limit, resulting in little or no amplitude increase from two to four items. In contrast, a subject with a high capacity of 4.5 items would be expected to be well below the limit for a two-item array, and should therefore show a large increase in amplitude for a four-item array.

The magnitude of the amplitude increase between two and four items was plotted as a function of each subject's memory capacity in Fig. 3b. These two measures were very strongly correlated ($r = 0.78$; $P < 0.0001$), with low-capacity subjects producing very little amplitude increase and high-capacity subjects showing larger amplitude increases. Importantly, an individual's memory capacity was not significantly correlated with either the amplitude increase between arrays of four and six items or the absolute amplitude of activity for a given array size.

These results show that the observed memory delay activity indexes the maintenance of active representations in visual memory. Moreover, they demonstrate a strong neurophysiological predictor of visual memory capacity. That is, simply by measuring the amplitude increase across memory array sizes, we could accurately predict an individual's memory capacity. Visual working memory is thought to have a central role within cognition because it maintains representations from the environment so that they may be acted on or manipulated^{3,26}. Indeed, an individual's ability to perform many

high-level cognitive functions has been shown to be directly influenced by his or her memory capacity^{5,27–29}. These results provide the first link between this important cognitive limitation and neural activity. □

Methods

Twelve neurologically normal college students participated in each experiment (age range of 21–33) and gave informed consent according to procedures approved by the University of Oregon. Each of these observers performed 240 trials per condition in each experiment. All stimulus arrays were presented within two $4^\circ \times 7.3^\circ$ rectangular regions that were centred 3° to the left and right of a central fixation cross on a grey background (8.2 cd m^{-2}). Each memory array consisted of 1–10 coloured squares ($0.65^\circ \times 0.65^\circ$) in each hemifield. Each square was selected at random from a set of seven highly discriminable colours (red, blue, violet, green, yellow, black and white), and a given colour could appear no more than twice within an array. Stimulus positions were randomized on each trial, with the constraint that the distance between squares within a hemifield was at least 2° (centre to centre). The colour of one square in the test array was different from the corresponding item in the memory array in 50% of trials; the colours of the two arrays were identical on the remaining trials. At the beginning of each trial, a central arrow cue instructed the subjects to remember the items in either the left or the right hemifield.

We computed visual memory capacity using a formula developed by Pashler²³ and refined by Cowan⁷. Essentially, this approach assumes that if an observer can hold K items in memory from an array of S items, then the item that changed should be one of the items being held in memory on K/S trials, leading to correct performance on K/S of the trials on which an item changed. To correct for guessing, this procedure also takes into account the false alarm rate. The formula is $K = S \times (H - F)$, where K is the memory capacity, S is the size of the array, H is the observed hit rate and F is the false alarm rate.

ERPs were recorded in each experiment using our standard recording and analysis procedures³⁰, including rejection of trials contaminated by blinks or large ($>1^\circ$) eye movements. We recorded from 22 standard electrode sites (international 10/20 system) spanning the scalp. We computed contralateral waveforms by averaging the activity recorded at right hemisphere electrode sites when subjects were cued to remember the left side of the memory array with the activity recorded from the left hemisphere electrode sites when they were cued to remember the right side. Contralateral delay activity was measured at posterior parietal, lateral occipital and posterior temporal electrode sites as the difference in mean amplitude between the ipsilateral and contralateral waveforms, with a measurement window of 300–900 ms after the onset of the memory array.

Received 23 December 2003; accepted 26 February 2004; doi:10.1038/nature02447.

1. Luck, S. J. & Vogel, E. K. The capacity of visual working memory for features and conjunctions. *Nature* **390**, 279–281 (1997).
2. Sperling, G. The information available in brief visual presentations. *Psychol. Monogr.* **74**, Whole No. 498 (1960).
3. Vogel, E. K., Woodman, G. F. & Luck, S. J. Storage of features, conjunctions, and objects in visual working memory. *J. Exp. Psychol. Hum. Percept. Perform.* **27**, 92–114 (2001).
4. Jevons, W. S. The power of numerical discrimination. *Nature* **3**, 281–282 (1871).
5. Engle, R. W., Kane, M. J. & Tuholski, S. W. In *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (eds Miyake, A. & Shah, P.) 102–134 (Cambridge Univ. Press, New York, 1999).
6. Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.* **63**, 81–97 (1956).
7. Cowan, N. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behav. Brain Sci.* **24**, 87–185 (2001).
8. Jonides, J. et al. Spatial working memory in humans as revealed by PET. *Nature* **363**, 623–625 (1993).
9. Courtney, S. M., Ungerleider, L. G., Keil, K. & Haxby, J. V. Transient and sustained activity in a distributed neural system for human working memory. *Nature* **386**, 608–611 (1997).
10. Cohen, J. D. et al. Temporal dynamics of brain activation during a working memory task. *Nature* **386**, 604–608 (1997).
11. Miller, E. K., Erickson, C. A. & Desimone, R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J. Neurosci.* **16**, 5154–5167 (1996).
12. Fuster, J. M. & Jervey, J. P. Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. *J. Neurosci.* **2**, 361–375 (1982).
13. Ruchkin, D., Johnson, R., Grafman, J., Canoune, H. & Ritter, W. Multiple visuospatial working memory buffers: Evidence from spatiotemporal patterns of brain activity. *Neuropsychologia* **35**, 195–209 (1997).
14. Ruchkin, D., Johnson, R., Canoune, H. & Ritter, W. Short-term memory storage and retention: An event-related brain potential study. *Electroencephalogr. Clin. Neurophysiol.* **76**, 419–439 (1990).
15. Gratton, G. The contralateral organization of visual processing: A theoretical concept and a research tool. *Psychophysiology* **35**, 638–647 (1998).
16. Woodman, G. F. & Luck, S. J. Electrophysiological measurement of rapid shifts of attention during visual search. *Nature* **400**, 867–869 (1999).
17. Miller, E. K., Li, L. & Desimone, R. Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *J. Neurosci.* **13**, 1460–1478 (1993).
18. Baddeley, A. Exploring the central executive. *Q. J. Exp. Psychol. A* **49**, 5–28 (1996).
19. Hillyard, S. A., Vogel, E. K. & Luck, S. J. Sensory gain control (amplification) as a mechanism of selective attention: Electrophysiological and neuroimaging evidence. *Phil. Trans. R. Soc. Lond. B* **353**, 1257–1270 (1998).
20. Rypma, B. & D'Esposito, M. D. The roles of prefrontal brain regions in components of working memory. *Proc. Natl Acad. Sci. USA* **96**, 6558–6563 (1999).
21. Ruchkin, D., Canoune, H., Johnson, R. & Ritter, W. Working memory and preparation elicit different patterns of slow wave event-related brain potentials. *Psychophysiology* **32**, 399–410 (1995).

22. Rypma, B. & D'Esposito, M. D. A subsequent-memory effect in dorsolateral prefrontal cortex. *Cogn. Brain Res.* **16**, 162–166 (2003).
23. Rypma, B., Prabhakaran, V., Desmond, J., Glover, G. H. & Gabrieli, J. D. Load-dependent roles of frontal brain regions in the maintenance of working memory. *Neuroimage* **9**, 216–226 (1999).
24. Eriksen, C. W. & St James, J. D. Visual attention within and around the field of focal attention: A zoom lens model. *Percept. Psychophys.* **40**, 225–240 (1986).
25. Pashler, H. Familiarity and visual change detection. *Percept. Psychophys.* **44**, 369–378 (1988).
26. Logie, R. H. *Visuo-Spatial Working Memory* (Erlbaum, Hove, UK, 1995).
27. Kane, M. J. & Engle, R. W. Working memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *J. Exp. Psychol. Gen.* **132**, 47–70 (2003).
28. Daneman, M. & Merikle, P. M. Working memory and language comprehension: A meta-analysis. *Psychon. Bull. Rev.* **3**, 422–433 (1996).
29. Kyllonen, P. C. & Christal, R. E. Reasoning ability is (little more than) working memory capacity. *Intelligence* **14**, 398–433 (1990).
30. Vogel, E. K., Luck, S. J. & Shapiro, K. L. Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. *J. Exp. Psychol. Hum. Percept. Perform.* **24**, 1656–1674 (1998).

Acknowledgements The research reported here was supported by a grant from the US National Institute of Mental Health.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.K.V. (vogel@darkwing.uoregon.edu).

Capacity limit of visual short-term memory in human posterior parietal cortex

J. Jay Todd & René Marois

Vanderbilt Vision Research Center, Department of Psychology, Vanderbilt University, 530 Wilson Hall, Nashville, Tennessee 37203, USA

At any instant, our visual system allows us to perceive a rich and detailed visual world. Yet our internal, explicit representation of this visual world is extremely sparse: we can only hold in mind a minute fraction of the visual scene^{1,2}. These mental representations are stored in visual short-term memory (VSTM). Even though VSTM is essential for the execution of a wide array of perceptual and cognitive functions^{3–5}, and is supported by an extensive network of brain regions^{6–9}, its storage capacity is severely limited^{10–13}. With the use of functional magnetic resonance imaging, we show here that this capacity limit is neurally reflected in one node of this network: activity in the posterior parietal cortex is tightly correlated with the limited amount of scene information that can be stored in VSTM. These results suggest that the posterior parietal cortex is a key neural locus of our impoverished mental representation of the visual world.

To investigate the neural basis of VSTM's storage capacity limit, 17 subjects were scanned while performing a parametric load manipulation¹⁴ of a delayed visual matching-to-sample task (Fig. 1). On each trial, subjects were briefly presented with a sample display containing one to eight coloured discs and, after a 1,200-ms retention interval, decided whether a single probe disc matched one of the sample discs in location and colour. A 1,200-ms delay maximizes VSTM's capacity: with delays shorter than 1 s, VSTM capacity is inflated by sensory (iconic) representations of the display¹⁵, whereas long delays not only underestimate VSTM capacity owing to memory degradation¹⁵, but also favour the recruitment of rehearsal mechanisms and verbal/abstract recoding of the visual material¹⁶. To minimize verbal strategies further, a verbal working-

memory/articulatory suppression task was administered concurrently with the VSTM task: throughout the trial, subjects rehearsed two digits presented at trial onset and reported them at trial offset. Performance in this task was high and independent of VSTM set size (92–94% accuracy across set sizes; $F_{5,80} = 0.64$, $P = 0.67$), attesting to the absence of a trade-off between the verbal and visual tasks, as predicted from the independence of these two working-memory systems^{17,18}.

Accuracy in the VSTM task declined with increased set size (set size 1, 97.7%; set size 2, 94.2%; set size 3, 90.0%; set size 4, 86.2%; set size 6, 73.3%; set size 8, 68.5%). The number of objects encoded at each set size, estimated with Cowan's K formula¹¹, increased up to set size 3 or 4, and levelled off thereafter (Fig. 2; t -test between set sizes 4 and 8, $P > 0.05$). This behavioural function is fitted significantly better by a quadratic function than by a linear function ($P = 0.01$)¹⁹. Thus, VSTM storage capacity is about three or four items, which is consistent with previous studies^{11,13}. Importantly, this capacity limit is not due to insufficient time to encode items in VSTM⁴. Tripling the sample presentation time from 150 to 450 ms in a separate experiment did not affect the K function ($n = 16$, $P = 0.28$), an observation consistent with previous findings^{12,13}. The VSTM task therefore expresses the capacity limit of VSTM storage as opposed to a limitation in spatially attending to the display or encoding items in VSTM.

The brain substrates mediating VSTM's storage capacity limit should demonstrate a response profile paralleling the behavioural K function: activation should increase until set size 3 or 4 and level off thereafter. To isolate such regions, a voxel-based multiple regression analysis with K -weighted set size coefficients was performed. The resulting statistical parametric maps revealed a single bilaterally symmetric area in the intraparietal and intraoccipital sulci (IPS/IOS; $P < 0.05$ corrected). Time-course analysis (Fig. 3a) confirmed a strong correlation between the IPS/IOS peak response amplitude and the number of objects encoded ($r = 0.54$, $P < 0.001$; Fig. 2). The peak blood oxygenation level-dependent (BOLD) response function reached a plateau by set size 4 (t -test between 4 and 8, $P < 0.05$) and was better described by a quadratic function than by a linear function ($P < 0.01$). This parietal activation is not simply related to task difficulty: accuracy decreased and reaction

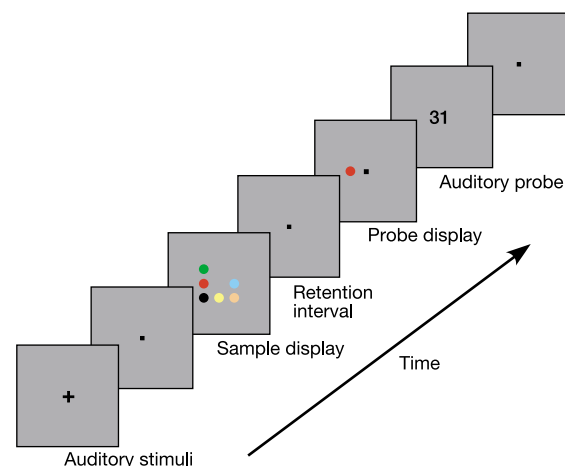


Figure 1 Trial design. Each trial began with the auditory presentation of two digits to be rehearsed throughout the trial. A sample display containing a variable number of coloured discs was then presented for 150 ms, followed by a 1,200-ms retention period, and then by a single coloured probe disc. Subjects judged whether the colour of the probe matched the colour of the disc shown at the same position in the sample display. Afterwards, two digits appeared and subjects indicated whether these were the same as those presented at trial onset.

time increased between set sizes 4 and 8 (P values < 0.05), when IPS/IOS activity and number of objects encoded remained constant. The levelling off of the BOLD signal at 0.3% is also unlikely to have resulted from haemodynamic saturation of the neurovascular system. In four subjects, we observed BOLD responses of more than 0.4% relative to fixation baseline in the same region of interest (ROI) when a working-memory task was presented at a short inter-trial interval (3 s), a condition that increases signal amplitude because of the summation of overlapping individual BOLD responses. Finally, the IPS/IOS involvement in VSTM generalizes across stimulus classes: it was also observed with white bars of varied orientations (0° , 45° , 90° , 135° and 180°) as stimuli (correlation of percentage signal change with K function of 0.64, $P < 0.01$, $n = 4$).

Could IPS/IOS activity reflect the perceptual or iconic representation of the number of objects in the scene instead of the number of objects stored in VSTM? To address this issue, we performed an iconic memory (IM) experiment that was identical to the first one except that for each trial only the sample display was presented, and subjects made an immediate judgement about the presence or absence of a coloured disc in the central position of the display. This task required attending to the display—with all stimuli presented at the fovea—and executing a response, but placed minimal and equal VSTM demands across set sizes. Performance was near ceiling across set sizes (96–98%). Relative to the VSTM experiment, the IPS/IOS BOLD response was attenuated and the set size effect was eliminated ($F_{1,5} = 1.13$, $P = 0.37$; Figs 2 and 3a). Thus, the IPS/IOS is insensitive to the perceptual load of the visual scene.

Visual working-memory tasks are traditionally decomposed into encoding, maintenance and retrieval phases^{14,20,21}. If the IPS/IOS is involved in VSTM storage, it should be engaged during maintenance, and not just at encoding or retrieval. Because of the short retention interval used in the VSTM experiment, the IPS/IOS response in each of these phases could not be distinguished. We therefore performed an additional VSTM load experiment with a sufficiently long retention interval (9,200 ms) to dissociate activity related to encoding, maintenance and retrieval^{21,22}. Only two set sizes were used to compensate for the lower number of trials acquired in this slow-event paradigm. As expected, the number of objects encoded at set size 3 was higher than at set size 1 ($K = 2.12$ versus 0.91, $P < 0.05$). Correspondingly, the IPS/IOS ROI was more activated at the larger set size during both encoding and maintenance (P values < 0.05 , one-tailed) but not during retrieval

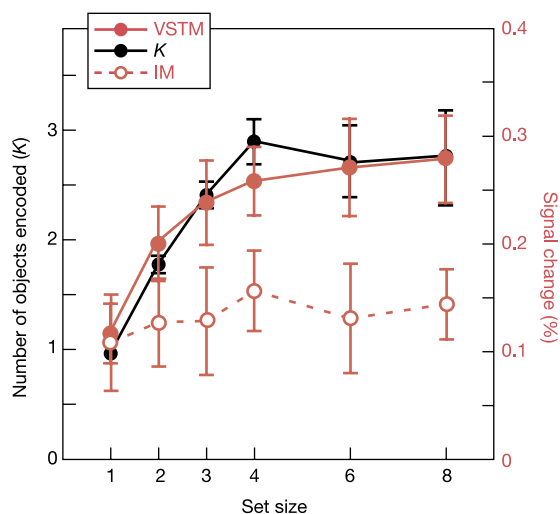


Figure 2 Behavioural performance and IPS/IOS response functions in VSTM and IM experiments. Behavioural performance corresponds to the estimated number (K) of encoded coloured discs at each set size. IPS/IOS BOLD signal response was measured at peak 5.6 s after stimulus presentation.

($P > 0.05$; Fig. 4). Thus, the IPS/IOS is sensitive to working-memory load chiefly during encoding and maintenance.

Because building a mental representation of a visual scene involves processing object identity (for example colour) and location, our VSTM task required encoding both visual attributes. Given the purported role of the posterior parietal cortex in visual feature integration^{23,24}, it is strategically positioned to build such integrated scene representations. However, ample evidence indicates that object identity and location are preferentially processed in ventral (occipito-temporal) and dorsal (parietal) cortical visual streams, respectively²⁵. It is therefore possible that at least partly distinct neural substrates underlie object and spatial VSTM capacity limits. An additional experiment investigated whether the IPS/IOS would still be involved in VSTM when only a judgement of object identity is required. This task was identical to the first experiment except that the location information was rendered task-irrelevant by always presenting the probe disc at fixation. Activation in the IPS/IOS ROI was still load-dependent ($F_{5,15} = 3.72$, $P < 0.05$, $n = 4$) and was correlated with the K function ($r = 0.70$, $t_3 = 3.67$, $P < 0.05$), indicating that the IPS/IOS might be involved in several forms of VSTM storage.

Do other brain regions besides IPS/IOS track VSTM storage capacity? To address this issue, we analysed the BOLD response in the two additional brain regions that emerged when the statistical parametric threshold was relaxed tenfold in the initial functional magnetic resonance imaging (fMRI) experiment (Fig. 3b, c): the anterior cingulate (AC) cortex and a region of ventral-occipital (VO) cortex with Talairach coordinates (Fig. 3) consistent with those of V4, an extrastriate area involved in colour processing²⁶. The

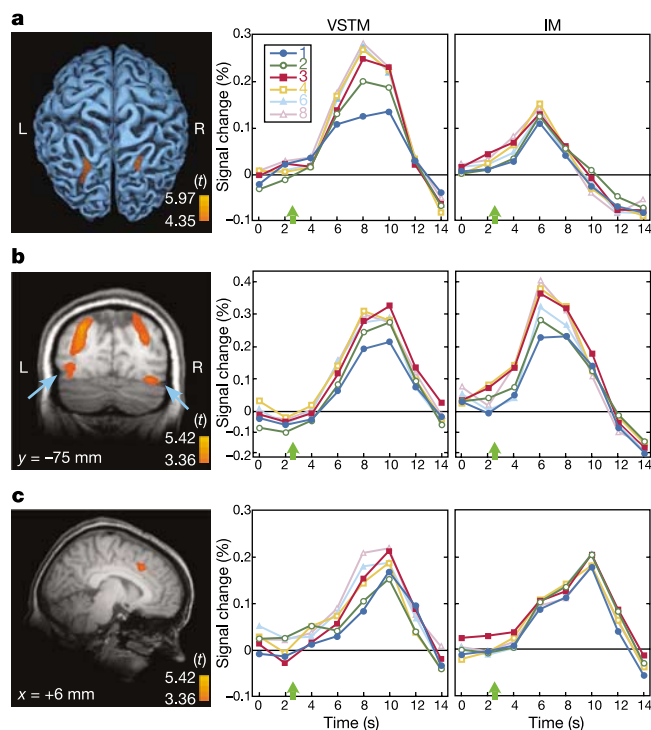


Figure 3 Brain activation time courses. Left, statistical parametric maps overlaid on structural scans of a representative subject. Middle and right, time courses for each set size in the VSTM and IM experiments, respectively. **a**, Bilateral IPS/IOS. Talairach coordinates (x , y , z): right/left, +23/-22 -59/-65 +45/+42. **b**, Bilateral VO ROI (blue arrows) with IPS/IOS activation dorsally. Right/left, +32/-34 -77/-68 -9/0. **c**, AC ROI, +8 +19 +30. The green arrow indicates the time of presentation of the visual stimulus. Although the activation baselines were variable across set sizes for VO and AC, the same statistical results are obtained when all time courses are further normalized to 0 at stimulus onset.

peak BOLD response functions of these two regions were distinct from the IPS/IOS (AC, $P < 0.001$; VO, $P < 0.05$), but not from a linear function (AC, $P = 0.40$; VO, $P = 0.13$). Additionally, neither AC nor VO showed a sustained response during maintenance in the extended VSTM retention experiment ($P > 0.05$; Fig. 4b, c), and both regions responded equally strongly under IM and VSTM conditions (P values > 0.05 ; Fig. 3b, c), indicating that their VSTM-related activations might be largely accounted for by perceptual and/or response components of the task. This is especially true for VO cortex, which, in stark contrast to the IPS/IOS, still showed a strong set size effect under IM conditions alone ($F_{1,5} = 4.48$, $P < 0.005$), indicating that ventral visual cortex might be driven by the perceptual load of the scene rather than by the amount of information that can be held in VSTM about the scene.

Although our results provide little evidence that the capacity limit of VSTM storage is a distributed property of the working-

memory network^{6,7,21}, they do not exclude the possibility of additional neural contributions below the level of fMRI detection, nor are they inconsistent with the established role of the frontal/prefrontal cortex in working memory^{14,20,21}. Whereas the posterior parietal cortex might act as a capacity-limited store for the representation of the visual scene, the frontal/prefrontal cortex might be necessary for the consolidation and/or maintenance of this store^{18,27}, especially during extended retention intervals.

Whatever the contributions of other brain regions might be, this series of experiments points to the posterior parietal cortex as a key neural locus of our impoverished mental representation of the visual world. Consistent with this notion, intraparietal cortex activity correlates with successful detection of a change between visual scenes²⁸, and the response of neurons in this area depends highly on the task relevance or saliency of stimuli in a multi-object display, a result that has led to the suggestion that scenes are sparsely represented in posterior parietal cortex²⁹. It is tempting to speculate that, because of the severe storage capacity limit of VSTM, the posterior parietal cortex's representation of the visual world could be nothing else but sparse. □

Methods

Subjects

Seventeen right-handed young adults (nine females) with normal or corrected-to-normal vision participated for financial compensation.

Task design

For the fast event-related fMRI experiment, each 8-s trial began with the auditory presentation, for 250 ms each, of the two digits to be rehearsed, followed by a 250-ms blank interval and a 250-ms auditory mask. At 1,500 ms after the auditory mask, the VSTM task began with presentation of the sample display for 150 ms (to preclude eye movements), with the display containing one, two, three, four, six or eight 0.38° discs, each of a different colour (either white, black, dark blue, light blue, orange, yellow, red, pink, dark green or light green), distributed among nine possible locations in a 3 × 3 matrix (1.38° × 1.38°). After a 1,200-ms retention period, a probe-coloured disc appeared for 1,750 ms at one of the occupied positions in the sample display. Subjects indicated by button press whether the probe's colour matched that of the sample. Half the trials were matched. For the other (unmatched) half, the probe colour was selected with equal probability from the other colours in the sample display or from the remaining colour set. After the visual probe response, two digits appeared for 1,500 ms and subjects indicated by button press whether the digits were identical to those rehearsed. Each fMRI run included seven iterations of each of the seven trial types (six set sizes and a non-event trial with no visual or auditory stimuli presented), with the order of trial type counterbalanced³⁰ within runs (seven runs per session).

For the slow event-related fMRI experiment. Six right-handed subjects (three females) from the pool of 17 subjects performed an additional experiment that was identical to the first except that the retention interval was extended from 1,200 to 9,200 ms (trial duration 18 s; seven trials per fMRI run) and trial types were randomly intermixed.

fMRI methods

Anatomical low-resolution and three-dimensional high-resolution T1-weighted images were acquired with conventional parameters on a 3-T GE MRI system (Milwaukee, Wisconsin, USA). Nineteen 7-mm-thick (3.75 mm × 3.75 mm in-plane, 0 mm skip) axial slices were taken parallel to the AC–PC line. T2*-weighted image parameters: echo time 25 ms; flip angle 70°; field of view 24 cm; 64 × 64 matrix; repetition time 2,000 ms. Trials were triggered by scanner pulses and presented with PsychToolBox for Matlab on an Apple G4 Macintosh. Stimuli were back-projected from an LCD projector on to a screen viewed through a prism mirror by the supine subject in the MR scanner.

Data analysis

For behavioural data, the number of objects encoded was estimated by using Cowan's K formula¹¹: for each set size, $K = (\text{hit rate} + \text{correct rejection rate} - 1)N$, where K is the number of objects encoded and N is the number of objects presented. This formula yields similar estimates to those obtained by related formulae^{12,13} and by full report (R.M., J.J.T. and C. M. Gilbert, unpublished observations).

For fMRI data, image analysis was performed with BrainVoyager 4.9.1 (Brain Innovation, Maastricht, The Netherlands). Data preprocessing included image realignment, three-dimensional motion correction, linear detrending, correction for slice scan acquisition order, and spatial smoothing with an 8-mm gaussian kernel (full width at half-maximum). Statistical parametric maps of BOLD activation for the fast event-related experiment were created by using a multiple regression analysis, with regressors defined for each trial type and convolved with a canonical haemodynamic function. The regression coefficients for each set size were weighted by their respective K estimates (balanced contrast design). The resulting maps from all subjects were standardized into Talairach space and superimposed to create cluster-filtered (equivalent of six contiguous 100-mm³ voxels in original space) composite maps. The overall model fit was assessed with an F value, and the obtained P values were corrected for the number of comparisons

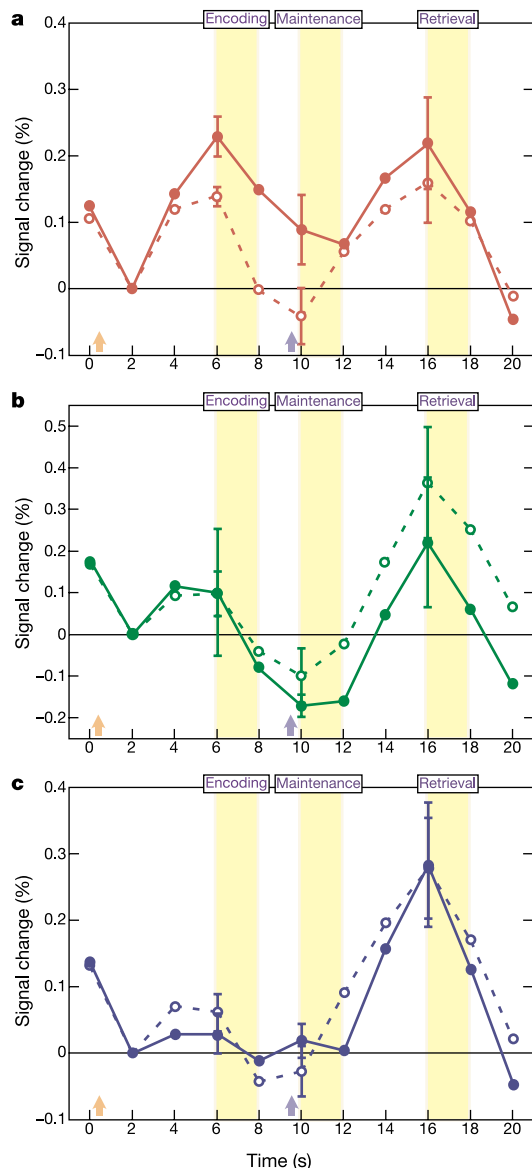


Figure 4 Response time courses during the encoding, maintenance and retrieval phases of a VSTM task with extended retention interval (9,200 ms). **a**, IPS/IOS; **b**, VO; **c**, AC. One (open circles) or three (filled circles) coloured discs were presented 1.6 s (orange arrow) before the standardized volume (time = 0 s). The purple arrow indicates the presentation of the probe display.

(<http://afni.nimh.nih.gov/afni/>) with a random-effects model. Time courses of each of the six set size conditions were extracted from each ROI to compute their respective percentage signal changes, with the no-event time series as baseline³⁰, first within each subject and subsequently averaged across subjects. The peak percentage signal change for each ROI was derived by collapsing the time courses of all set sizes and subjects and determining the time point of greatest signal amplitude in the grand average. One subject with unsteady raw signal who failed to show any time-locked response was removed from further analysis. The slow event-related experiment was analysed as above, except that the percentage signal change of each time course was normalized to the first image acquired after stimulus presentation (baseline), because the signal from the previous image still included the falling phase of the response to the previous trial (Fig. 4). The peaks of the encoding, maintenance and retrieval phases corresponded to the volumes acquired 5.5–7.5, 9.5–11.5 and 15.5–17.5 s after stimulus presentation²¹, respectively.

Received 11 January; accepted 9 March 2004; doi:10.1038/nature02466.

1. Rensink, R. A. Change detection. *Annu. Rev. Psychol.* **53**, 245–277 (2002).
2. Simons, D. & Levin, D. Change blindness. *Trends Cogn. Sci.* **1**, 261–267 (1997).
3. Chun, M. M. & Potter, M. C. A two-stage model for multiple target detection in rapid serial visual presentation. *J. Exp. Psychol. Hum. Percept. Perform.* **21**, 109–127 (1995).
4. Jolicoeur, P., Dell'Acqua, R. & Crebolder, J. M. in *The Limits of Attention: Temporal Constraints in Human Information Processing* (ed. Shapiro, K.) 82–99 (Oxford Univ. Press, 2001).
5. Wheeler, M. E. & Treisman, A. M. Binding in short-term visual memory. *J. Exp. Psychol. Gen.* **131**, 48–64 (2002).
6. Goldman-Rakic, P. S. in *Handbook of Physiology: The Nervous System, Higher Functions of the Brain* (eds Mountcastle, V. B. & Plum, F.) 373–417 (American Physiological Society, Bethesda, Maryland, 1987).
7. Ungerleider, L. G., Courtney, S. M. & Haxby, J. V. A neural system for human visual working memory. *Proc. Natl Acad. Sci. USA* **95**, 883–890 (1998).
8. Callicott, J. H. et al. Physiological characteristics of capacity constraints in working memory as revealed by functional MRI. *Cereb. Cortex* **9**, 20–26 (1999).
9. Linden, D. E. et al. Cortical capacity constraints for visual working memory: Dissociation of fMRI load effects in a fronto-parietal network. *Neuroimage* **20**, 1518–1530 (2003).
10. Duncan, J. et al. Systematic analysis of deficits in visual attention. *J. Exp. Psychol. Gen.* **128**, 450–478 (1999).
11. Cowan, N. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behav. Brain Sci.* **24**, 87–114 (2001).
12. Pashler, H. Familiarity and visual change detection. *Percept. Psychophys.* **44**, 369–378 (1988).
13. Vogel, E. K., Woodman, G. F. & Luck, S. J. Storage of features, conjunctions and objects in visual working memory. *J. Exp. Psychol. Hum. Percept. Perform.* **27**, 92–114 (2001).
14. Cohen, J. D. et al. Temporal dynamics of brain activation during a working memory task. *Nature* **386**, 604–608 (1997).
15. Phillips, W. A. On the distinction between sensory storage and short-term visual memory. *Percept. Psychophys.* **16**, 283–290 (1974).
16. Coltheart, M. in *New Horizons in Psychology* (ed. Dodwell, P. C.) 62–85 (Harmondsworth, Penguin, 1972).
17. Baddeley, A. Working memory. *Science* **255**, 556–559 (1992).
18. Smith, E. E. & Jonides, J. Neuroimaging analyses of human working memory. *Proc. Natl Acad. Sci. USA* **95**, 12061–12068 (1998).
19. Rosnow, R. L. & Rosenthal, R. Contrasts and interactions redux: Five easy pieces. *Psychol. Sci.* **7**, 253–257 (1996).
20. Courtney, S. M., Ungerleider, L. G., Keil, K. & Haxby, J. V. Transient and sustained activity in a distributed neural system for human working memory. *Nature* **386**, 608–611 (1997).
21. Pessoa, K., Gutierrez, E., Bandettini, P. A. & Ungerleider, L. G. Neural correlates of visual working memory: fMRI amplitude predicts task performance. *Neuron* **35**, 975–987 (2002).
22. Zarahn, E., Aguirre, G. & D'Esposito, M. A trial-based experimental design for fMRI. *Neuroimage* **6**, 122–138 (1997).
23. Friedman-Hill, S. R., Robertson, L. C. & Treisman, A. Parietal contributions to visual feature binding: Evidence from a patient with bilateral lesions. *Science* **269**, 853–855 (1995).
24. Shafritz, K. M., Gore, J. C. & Marois, R. The role of the parietal cortex in visual feature binding. *Proc. Natl Acad. Sci. USA* **99**, 10917–10922 (2002).
25. Haxby, J. V. et al. Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proc. Natl Acad. Sci. USA* **88**, 1621–1625 (1991).
26. McKeefry, D. J. & Zeki, S. The position and topography of the human colour centre as revealed by functional magnetic resonance imaging. *Brain* **120**, 2229–2242 (1997).
27. Curtis, C. E. & D'Esposito, M. Persistent activity in the prefrontal cortex during working memory. *Trends Cogn. Sci.* **7**, 415–423 (2003).
28. Beck, D. M., Rees, G., Frith, C. D. & Lavie, N. Neural correlates of change detection and change blindness. *Nature Neurosci.* **4**, 645–650 (2001).
29. Gottlieb, J. P., Kusunoki, M. & Goldberg, M. E. The representation of visual salience in monkey parietal cortex. *Nature* **391**, 481–484 (1998).
30. Kourtzi, Z. & Kanwisher, N. Representation of perceived object shape by the human lateral occipital complex. *Science* **293**, 1506–1509 (2001).

Acknowledgements We thank I. Gauthier, M. Chun, G. Logan and J. Schall for comments on earlier versions of this manuscript, and D. Nikolaiczky-Stocks and A. Snyder for expert technical assistance. This work was supported by a grant from the NSF to R.M.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.M. (rene.marois@vanderbilt.edu).

The endothelial-cell-derived secreted factor *Egfl7* regulates vascular tube formation

Leon H. Parker¹, Maike Schmidt¹, Suk-Won Jin³, Alane M. Gray¹, Dimitris Beis³, Thinh Pham², Gretchen Frantz², Susan Palmieri², Kenneth Hillan², Didier Y. R. Stainier³, Frederic J. de Sauvage¹ & Weilan Ye¹

¹Molecular Biology Department and ²Pathology Department, Genentech Inc., South San Francisco, California 94080, USA

³Department of Biochemistry and Biophysics, Programs in Developmental Biology, Genetics and Human Genetics, University of California San Francisco, San Francisco, California 94143-0448, USA

Vascular development is a complex but orderly process that is tightly regulated. A number of secreted factors produced by surrounding cells regulate endothelial cell (EC) differentiation, proliferation, migration and coalescence into cord-like structures^{1,2}. Vascular cords then undergo tubulogenesis to form vessels with a central lumen^{3,4}. But little is known about how tubulogenesis is regulated *in vivo*. Here we report the identification and characterization of a new EC-derived secreted factor, *EGF-like domain 7 (Egfl7)*. *Egfl7* is expressed at high levels in the vasculature associated with tissue proliferation, and is downregulated in most of the mature vessels in normal adult tissues. Loss of *Egfl7* function in zebrafish embryos specifically blocks vascular tubulogenesis. We uncover a dynamic process during which gradual separation and proper spatial arrangement of the angioblasts allow subsequent assembly of vascular tubes. This process fails to take place in *Egfl7* knockdown embryos, leading to the failure of vascular tube formation. Our study defines a regulator that controls a specific and important step in vasculogenesis.

Tubulogenesis is an essential step in vascular development. Vascular tube formation proceeds through two phases: first, ECs coalesce to form cords after they reach their destination; cords are then reshaped into tubes^{3,4}. But we have little understanding of how vascular cords progress to become tubes and what factors regulate this transition *in vivo*. In an effort to identify molecules involved in the regulation of vascular development, we examined the expression pattern of more than 600 new secreted and transmembrane factors by *in situ* hybridization in whole mouse embryos ranging from embryonic day 7.5 (E7.5) to E14.5, a time window encompassing many key steps in vasculogenesis (vessel formation from EC progenitors) and angiogenesis (sprouting from existing vessels)^{1,2,5}. One of the factors we identified in this screen is a secreted factor known as *EGF-like domain 7 (Egfl7)*, also called *zneul*, *neul*, *notch4-like*, *tango125*, *VE-statin*.

Egfl7 encodes a putative secreted protein with a relative molecular mass of ~30 K that is evolutionarily conserved. The *EGFL7* protein contains a signal sequence, an EMI domain at the amino terminus (the EMI domain is present in a number of extracellular-matrix-associated proteins involved in regulating cell adhesion^{6,7}), followed by two EGF-like domains and a leucine- and valine-rich carboxy-terminal region (Supplementary Fig. 1a). The mammalian *Egfl7* belongs to a small gene family. Through BLAST searches, we identified one closely related gene, *Egfl8*, which has a domain organization identical to that of *Egfl7*. Interestingly, this gene family seems to be more complex in mammals, as we have not been able to identify an *Egfl8* orthologue in several fish genomes (*Danio rerio*, *Medaka* and *Fugu*) (Supplementary Fig. 1c).

The expression pattern of *Egfl7* is conserved across species. In mouse, human and zebrafish embryos, high levels of *Egfl7* transcripts are detected in endothelial progenitors and ECs in all vessels

(Fig. 1a, b, d and g, j–m; Supplementary Fig. 2) as well as the endocardium (data not shown). Strong vascular expression persists throughout embryonic and neonatal development (data not shown); however, the message is undetectable in many adult organs (Fig. 1e and h; Supplementary Fig. 2j). In adult mice, a few highly vascularized organs such as the lung, heart and kidney continue to express *Egfl7* in a small subset of vessels (Supplementary Fig. 2k). Interestingly, *Egfl7* expression is strongly upregulated in many proliferative tissues, including tumours (Fig. 1f and i; Supplementary Fig. 2f), reproductive organs during pregnancy (Fig. 1c; Supplementary Fig. 2d) and inflamed tissues (data not shown). This unique expression pattern suggests that high levels of *Egfl7* are associated with vascular growth and remodelling. *Egfl7* is not detected outside the cardiovascular and haematopoietic systems; this observation is confirmed by the fact that *Egfl7* expression is

completely abolished in the avascular zebrafish mutant *cloche*⁸ (Fig. 1n). Immunofluorescence staining shows that the EGFL7 protein is secreted but remains in the vicinity of the ECs (Fig. 1c, inset), possibly associated with the extracellular matrix.

In a recent report, conditioned medium containing the recombinant EGFL7 protein is shown to inhibit smooth muscle cell (SMC) migration *in vitro*⁹. However, its *in vivo* function has not been defined. To uncover the *in vivo* biological function, we performed gene knockdown experiments using two different morpholino antisense oligonucleotides¹⁰ targeting the zebrafish *Egfl7*: oligonucleotide AS₄₇ hybridizes to the 5' untranslated region (UTR) and blocks translation; oligonucleotide AS₁₉₅ hybridizes with an exon–intron junction, resulting in intron retention and hence premature translation termination (Supplementary Figs 1b and 3e). Both oligonucleotides give identical results. When examined

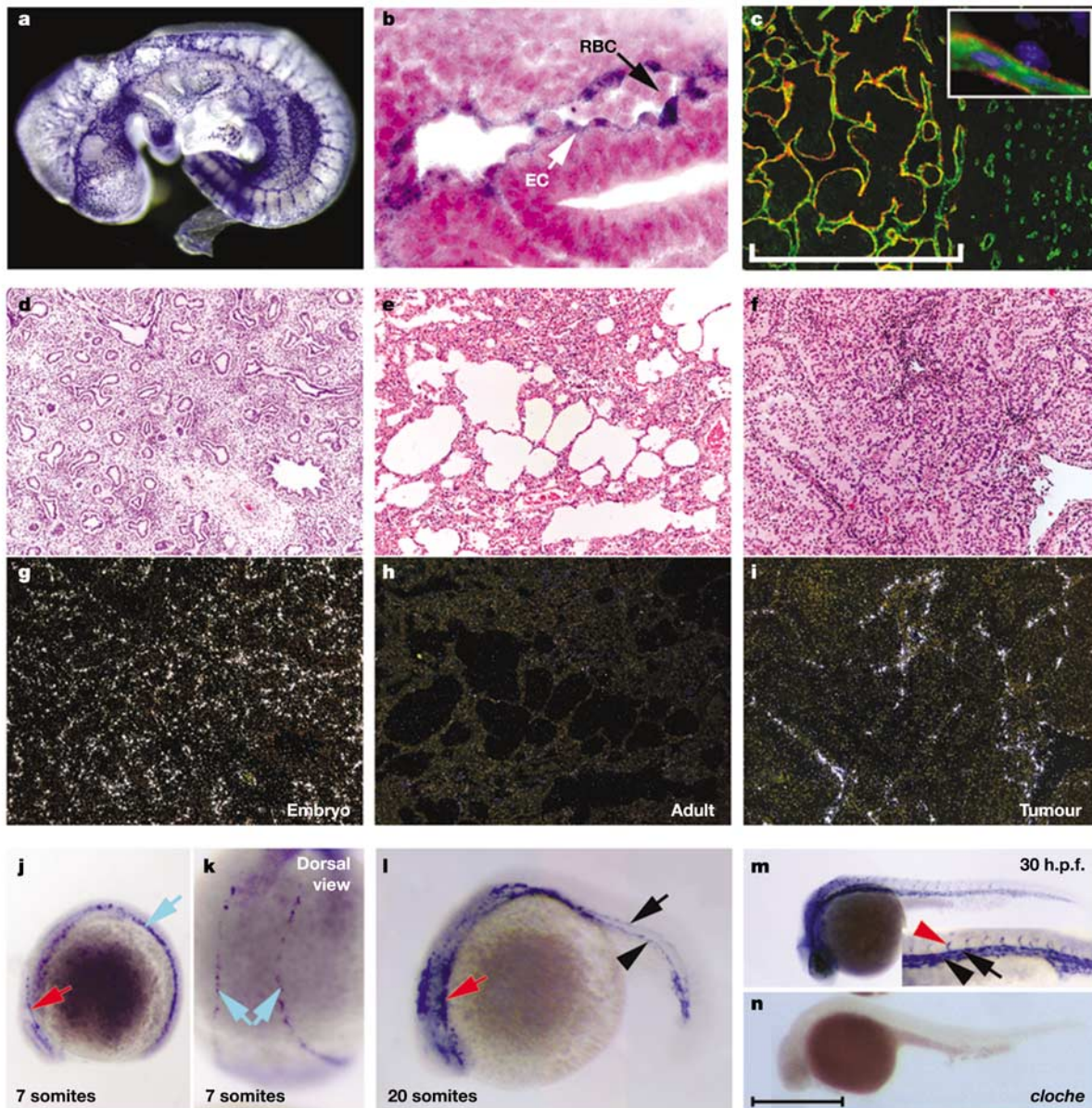


Figure 1 *Egfl7* expression profile. *Egfl7* wholemount *in situ* hybridization on mouse (a, b) and zebrafish (j–n) embryos. a, E9 whole embryo. b, Cross-section of a stained with nuclear fast red. RBC, red blood cells. j–m, Blue arrow, lateral plate mesoderm; red arrow, head vascular primordia; black arrow, dorsal aorta; black arrowhead, axial vein; red arrowhead, ISVs. Inset, close-up of the trunk. n, *cloche* mutant. c, Pregnant mouse

uterus stained for EGFL7 (red) and platelet/EC adhesion molecule 1 (PECAM-1) (green). Bracket, decidua. d–i, Radioactive *in situ* hybridization (g–i) and haematoxylin and eosin (H&E) staining (d–f) on human lung sections. Scale bar, 0.45 mm (a, m, n), 0.07 mm (b), 0.38 mm (c–i), 0.25 mm (j, l), 0.15 mm (k), 0.26 mm (m, inset) and 0.04 mm (c, inset).

after the onset of circulation, ~24 hours post fertilization (h.p.f.), more than 40% of the *Egfl7* knockdown embryos (KDs) show overt signs of vascular defects: they either have no circulation at all or develop an incomplete circulatory loop (Supplementary Movies), and many have pericardial oedema and haemorrhage (Fig. 2a). In contrast, only 3% of the control oligonucleotide-injected fish have minor vascular defects (Supplementary Fig. 3a).

We analysed the expression patterns of several vascular endothelial markers, including *fli1* (ref. 11), *flk1* (refs 12, 13), *tie1* (ref. 14), *ephrinB2* and *gridlock* (arterial EC marker), *flt4* and *EphB4* (venous EC marker)^{15,16}, and endogenous alkaline phosphatase activity¹⁷. We also carried out knockdown experiments in an *flk1*-promoter-GFP (green fluorescent protein) transgenic line (D.B., J. N. Chen and D.Y.R.S., unpublished data). From the 7-somite stage to 30 h.p.f., the overall spatial distribution and intensity of the above markers are unaffected in the KDs; furthermore, all primary arteries and veins are formed in the correct locations in the KDs (Figs 2 and 3; Supplementary Figs 3b–d and 4f), indicating that there is no significant defect in early EC differentiation, proliferation and migration. However, tubulogenesis throughout the entire system is disrupted in the KDs. At 30 h.p.f., many primary vessels in the KDs have either disorganized lumens or no lumen at all (Figs 2c, d and 3h; Supplementary Fig. 3b–d). The prevalence of the tubulogenesis defect displayed by molecular markers is between 75% and 85% in multiple experiments. For instance, *fli1* staining at 30 h.p.f. reveals that 76% (28/37) of the KDs have a tubulogenesis defect, whereas all control embryos ($n = 37$) develop normal vascular tubes. To confirm the knockdown specificity, we show that the vascular phenotype caused by AS₋₄₇ is rescued by the *Egfl7* coding RNA without the 5' UTR (Supplementary Fig. 3f).

Lack of lumen formation in the *Egfl7* KDs is verified by the finding that major vessels in these embryos cannot be filled with dye (Supplementary Fig. 4a–d). Although it is possible that the tubulogenesis defect is a consequence of vascular collapse due to lack of

blood flow, we believe that this is unlikely on the basis of the following observations: first, cardiac contractile function is normal in the *Egfl7* KDs (Supplementary Movies); second, primary vessel lumen formation and short-term maintenance are normal in the *silent heart* mutants that lack circulation^{18,19} (Supplementary Fig. 4e).

We also observed progressive angiogenesis defects in the *Egfl7* KDs. The initial sprouting of intersegmental vessels (ISVs) occurs normally in the mutants at 22–24 h.p.f. (Fig. 2b). However, these secondary vessels gradually disappear, as the ISVs are partially missing at 30 h.p.f. (Fig. 2c, d) and are completely eliminated by 48–72 h.p.f. (Supplementary Fig. 3f). It is likely that a lack of circulation contributes to the angiogenesis defect, as it is documented that ISVs undergo regression in the absence of blood flow¹⁹. Owing to the dominant phenotype in the primary vessels, the exact role of *Egfl7* in angiogenesis remains to be defined using other methods, such as inducible knockout.

We then examined whether the vascular defect seen in the *Egfl7* KDs is an indirect consequence of disrupting other tissues, because vascular development is dependent on neighbouring tissues^{20,21}. In cross-sections of the KDs, we saw a normal complement of tissues (Fig. 3). Furthermore, *in situ* hybridization with *fkd7* (ref. 22), *ntl* (ref. 23), *axial/flk1* (ref. 24) and *gata1* (ref. 25) illustrates that all axial structures and the haematopoietic lineage develop normally in the KDs (Supplementary Fig. 4g–i and data not shown). After the onset of circulation, *gata1*⁺ cells in the KDs remain in the posterior

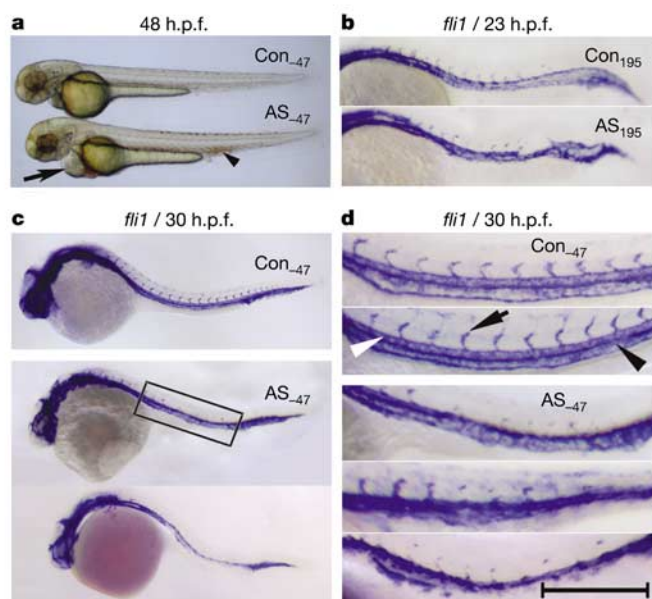


Figure 2 *Egfl7* gene knockdown causes a vascular tubulogenesis defect in zebrafish embryos. Zebrafish embryos injected with control (Con₋₄₇ or Con₁₉₅) or *Egfl7* antisense (AS₋₄₇ or AS₁₉₅) oligonucleotides. **a**, Gross morphology at 48 h.p.f. Arrow points to pericardial oedema; arrowhead indicates haemorrhage. **b–d**, *fli1* expression at 23 h.p.f. (**b**) and 30 h.p.f. (**c–d**). **d**, Close-up view of the mid-trunk vasculatures boxed in **c**. White arrowhead, lumen of the dorsal aorta; black arrowhead, lumen of the posterior cardinal vein; black arrow, intersegmental vessels. Scale bar, 0.6 mm (**a**), 0.23 mm (**d**) and 0.5 mm (**b**, **c**).

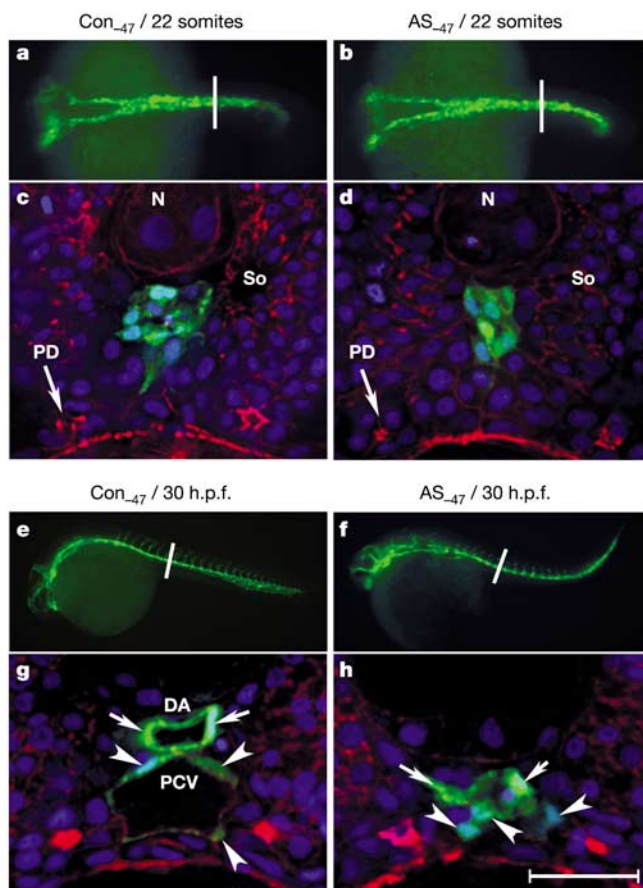


Figure 3 Endothelial cell number is unaltered in the *Egfl7* knockdowns. *flk1:GFP* transgenic fish injected with control (**a**, **c**, **e**, **g**) or antisense (**b**, **d**, **f**, **h**) oligonucleotides were analysed at the 22-somite stage (**a–d**) or at 30 h.p.f. (**e–h**). **a**, **b**, Dorsal view. **e**, **f**, Lateral view. **c**, **d**, **g**, **h**, Cross-sections taken at the level indicated by white lines in **a**, **b**, **e** and **f**, counter-stained with phalloidin (red) and 4,6-diamidino-2-phenylindole (DAPI; blue). PD, pronephric ducts; So, somites; N, notochords. **g**, **h**, White arrows, arterial ECs; white arrowheads, venous ECs; DA, dorsal aorta; PCV, posterior cardinal vein. Scale bar, 0.33 mm (**a**, **b**), 0.03 mm (**c**, **d**, **g**, **h**) and 0.47 mm (**e**, **f**).

ventrolateral region of the embryo, where they initially developed, confirming that the vasculature is defective. Finally, defective recruitment of SMC is an unlikely causative factor because we found no evidence of perivascular *SM22 α* expression in wild-type embryos at the 26-somite stage (data not shown), a time point at which the vascular defects are obvious in the KDs (Supplementary Fig. 5e–f). Taken together, our data favour the idea that failure of vascular tubulogenesis is a primary defect caused by *Egfl7* knockdown.

Because EC number dictates vascular morphogenesis²⁶, we investigated whether EC number is altered in the *Egfl7* KDs. Serial cross-sections show that knocking down *Egfl7* does not change the total number of ECs at any axial level, regardless of developmental stage (Fig. 3). Furthermore, on the basis of *ephrinB2* and *flt4* expression, and the differential regulation of *flk1* promoter between artery and vein in the *flk1:GFP* embryos (D.B., J. N. Chen and D.Y.R.S., unpublished data), we found that arterial and venous EC numbers

are also unaffected (Fig. 3 and data not shown). In contrast to *Egfl7*, knocking down *vegfr* reduces EC number significantly and, as an indirect consequence, disrupts tube formation (Supplementary Fig. 4j–l). Our data indicate that *Egfl7* is specifically required for vascular tubulogenesis, a role that is distinctive from that of *vegfr*.

To describe the vascular phenotype at a cellular level, we carried out a time-course analysis using the *flk1:GFP* fish. Serial cross-sections of the trunk and longitudinal sections of the head were taken at the 22-, 24- and 26-somite stages (before circulation), and at 24 h.p.f. (onset of circulation) and 30 h.p.f. Analyses of these sections reveal a series of cellular events during the cord-to-tube transition in the control embryos. At the 22-somite stage, arterial and venous angioblasts coalesce into single cords. Extensive tight junctions reveal intimate connections between angioblasts (Supplementary Fig. 5a). Around the 24-somite stage, gradual separation of angioblasts is evident by substantial refinement of tight junctions (Supplementary Fig. 5c). By the 26-somite stage, angioblasts are segregated such that arterial and venous angioblasts occupy distinct domains and align in the form of rudimentary tubes (Supplementary Fig. 5e). Subsequently, ECs undergo extensive morphological changes and become squamous, lending the vascular tubes their final shape (Fig. 3g; Supplementary Fig. 5g). This sequence of events leading to the formation of major vascular tubes is impaired in the *Egfl7* KDs. At all the stages examined, the knockdown angioblasts fail to separate and retain extensive tight junctions (Supplementary Fig. 5d, f). They also fail to change shape at the later stages (Fig. 3h; Supplementary Fig. 5h).

Failure of EC separation indicates that either EC motility or adhesion is improperly regulated in the *Egfl7* KDs. To distinguish between these two possibilities, we carried out *in vitro* EC migration and adhesion assays. We found that recombinant EGFL7 protein does not promote EC migration (data not shown); however, EGFL7 coated on culture plates promotes human umbilical vein endothelial cell (HUVEC) adhesion (Fig. 4). Interestingly, the strength of adhesion promoted by EGFL7 is significantly weaker than that of other classic cell-adhesion molecules, such as fibronectin and collagens (Fig. 4e). Moreover, the kinetics of HUVEC adhesion to EGFL7 is much slower than that of other substrates (Fig. 4g). Taking into account the EGFL7 expression pattern, subcellular localization, *in vivo* function and cell-adhesion properties, we propose that during active vascular growth, EGFL7 provides a permissive substrate that favours motility over stable attachment, thereby allowing the local movement of angioblasts that is required for tube formation (Supplementary Fig. 6).

In summary, we have shown that *Egfl7* has a specific role in regulating vascular tubulogenesis, and define a dynamic process during which endothelial cell–cell connections and local migration are coordinately controlled to ensure proper assembly of major vascular tubes. Understanding the mechanisms that regulate EGFL7 production and its signalling pathway should aid in advancing our knowledge of vascular development, and may enable us to design novel therapeutic strategies for the treatment of diseases associated with pathological angiogenesis. □

Methods

Zebrafish and mouse strains

Mouse embryos were harvested from timed-pregnant CD-1 mice from Charles River. Tübingen long fin (TL) wild-type zebrafish were used for microinjection. *cloche* (*clo*^{m39}) homozygous mutant embryos at 30 h.p.f. and their wild-type siblings were kindly provided by M. Fishman.

Microinjection of oligonucleotides and RNA

Morpholino (MO) antisense and control oligonucleotides (see Supplementary Information for sequences) were ordered from Sequitur, Inc. and GeneTools LLC. Oligonucleotide sequences are indicated in Supplementary Fig. 1. The MO results were further confirmed using PNA oligonucleotides from Active Motif Inc. Injection of 4 ng of antisense oligonucleotide per embryo caused specific vascular defects with minimal non-specific defects. 5'-capped zebrafish *Egfl7* RNA (0.5 ng) was co-injected with 4 ng AS₄₇ oligonucleotide per embryo in the rescue experiment.

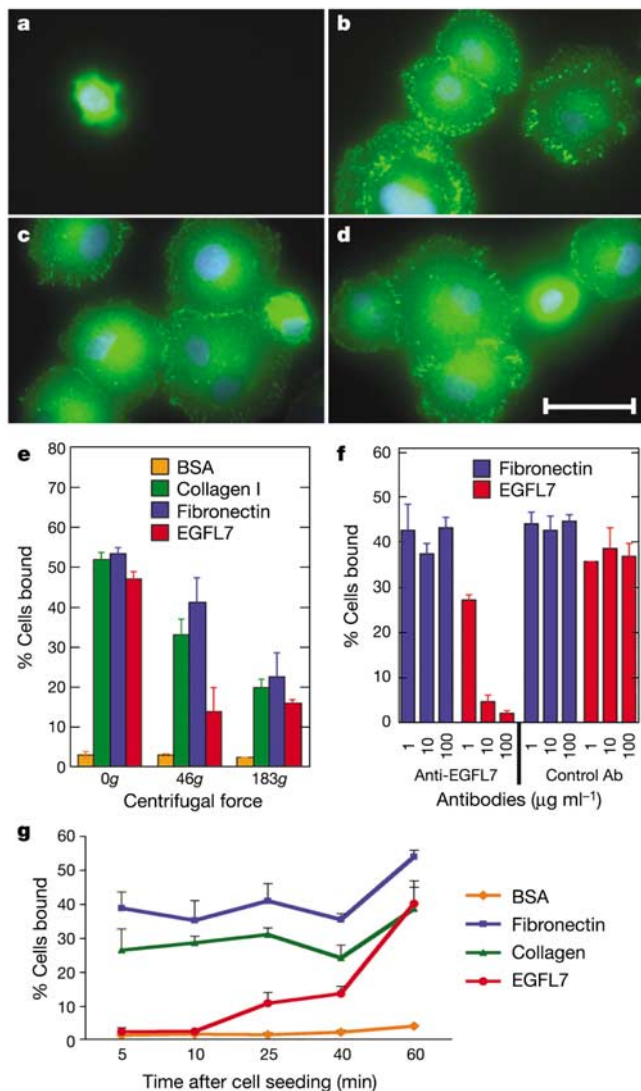


Figure 4 EGFL7 promotes endothelial cell adhesion. Vinculin staining of HUVECs reveals focal adhesion formation on fibronectin (**b**), type I collagen (**c**) and EGFL7 (**d**), but not BSA (**a**). Adhesion strength on EGFL7 is weaker than on collagen or fibronectin because fewer cells remain adherent to the EGFL7 substrate after centrifugation at 46g (**e**). Dose-dependent blockage of HUVEC adhesion to EGFL7 but not fibronectin by an anti-EGFL7 antibody confirms substrate specificity. A control antibody (anti-B7x) has no effect on any substrate (**f**). **g**, Kinetics of HUVEC adhesion on various substrates. Scale bar, 0.03 mm (**a–d**).

In situ hybridization

See Supplementary Fig. 2.

Alkaline phosphatase staining

Endogenous alkaline phosphatase activity in 48–72 h.p.f. wholemount zebrafish embryos was detected as described previously¹⁷.

Deconvolution microscopy

Embryos were fixed overnight at 4 °C in 4% PFA, cryoprotected in 20% sucrose and PBS, snap frozen in 7.5% gelatin and 15% sucrose, and sectioned. Slides were air dried, permeabilized in 1 × PBS, 0.1% Triton X-100 and 1% DMSO, blocked for 20 min in the above solution plus 1% BSA, stained with 66 nM ALEXA Fluor594-phalloidin (Molecular Probes) for 20 min, washed, mounted with Vectashield DAPI (Vector labs), and analysed using a Deltavision Deconvolution Microscope with a ×60 oil objective.

Immunofluorescent staining

Armenian hamsters were immunized with recombinant murine EGFL7 protein expressed in *Escherichia coli*. Monoclonal antibodies were generated by hybridoma fusion and subcloning. Two monoclonal antibodies, 1C8 and 5H7, that recognize different epitopes were used for immunofluorescent staining. Mouse tissues were snap frozen and cryosectioned. Fixed cells or 5 µm unfixed tissue sections were stained as described²⁷. Antibodies used in this study were: anti-ZO-1 monoclonal antibody (catalogue no. 33-9100, Zymed Inc.), anti-GFP (catalogue no. TP401, Torrey Pines Biolabs) and anti-vinculin monoclonal antibody (Sigma).

Cell-adhesion assay

Plates were coated with 5 µg cm⁻² protein (BSA (Sigma), collagen (Upstate), fibronectin (Sigma) and recombinant human EGFL7 produced in *E. coli* at Genentech). After PBS rinses, HUVECs (Cambrex) were plated at a density of 5 × 10³ cm⁻² in EGM2 medium (Cambrex) and centrifuged for 5 min at 140g to synchronize cell attachment, and then incubated. To analyse specificity, plates were pre-incubated with the indicated concentrations of antibody before HUVEC plating. Monoclonal anti-human EGFL7 antibody 1B12 was generated at Genentech using the recombinant human EGFL7 protein as immunogen. To determine adhesion strength, inverted plates were centrifuged at 46g, 183g or 411g after 60 min of incubation. The number of adherent cells was quantified using a fluorescence-based assay (CyQUANT, Molecular Probes) and readouts were taken using a fluorescence plate reader (Spectramax, Molecular Devices).

Accession numbers

GenBank accession numbers for *Egfl7* are as follows: NM_016215 (*Homo sapiens Egfl7/VE-statin*), NM_178444 (*Mus musculus Egfl7*), AF184973 (*M. musculus Notch4-like*), P_AA37135 (*M. musculus TANGO125*), BC044267 (*Xenopus laevis NEU1*) and AY542170 (*D. rerio Egfl7*). *Egfl8* accession numbers are: NM_030652 (*H. sapiens*) and NM_152922 (*M. musculus*).

Received 4 November 2003; accepted 13 February 2004; doi:10.1038/nature02416.

- Carmeliet, P. & Jain, R. K. Angiogenesis in cancer and other diseases. *Nature* **407**, 249–257 (2000).
- Hanahan, D. Signaling vascular morphogenesis and maintenance. *Science* **277**, 48–50 (1997).
- Hogan, B. L. & Kolodziej, P. A. Organogenesis: molecular mechanisms of tubulogenesis. *Nature Rev. Genet.* **3**, 513–523 (2002).
- Lubarsky, B. & Krasnow, M. A. Tube morphogenesis: making and shaping biological tubes. *Cell* **112**, 19–28 (2003).
- Folkman, J. Angiogenesis and angiogenesis inhibition: an overview. *EXS* **79**, 1–8 (1997).
- Doliana, R., Bot, S., Bonaldi, P. & Colombatti, A. EMI, a novel cysteine-rich domain of EMILINs and other extracellular proteins, interacts with the gC1q domains and participates in multimerization. *FEBS Lett.* **484**, 164–168 (2000).
- Callebaut, I., Mignotte, V., Souchet, M. & Mornon, J. P. EMI domains are widespread and reveal the probable orthologs of the *Caenorhabditis elegans* CED-1 protein. *Biochem. Biophys. Res. Commun.* **300**, 619–623 (2003).
- Stainier, D. Y., Weinstein, B. M., Detrich, H. W. III, L. I. & Fishman, M. C. *cloche*, an early acting zebrafish gene, is required by both the endothelial and hematopoietic lineages. *Dev. Suppl.* **121**, 3141–3150 (1995).
- Soncin, F. *et al.* VE-statin, an endothelial repressor of smooth muscle cell migration. *EMBO J.* **22**, 5700–5711 (2003).
- Nasevicius, A. & Ekker, S. C. Effective targeted gene 'knockdown' in zebrafish. *Nature Genet.* **26**, 216–220 (2000).
- Brown, L. A. *et al.* Insights into early vasculogenesis revealed by expression of the ETS-domain transcription factor Flt-1 in wild-type and mutant zebrafish embryos. *Mech. Dev.* **90**, 237–252 (2000).
- Fouquet, B., Weinstein, B. M., Serluca, F. C. & Fishman, M. C. Vessel patterning in the embryo of the zebrafish: guidance by notochord. *Dev. Biol.* **183**, 37–48 (1997).
- Liao, W. *et al.* The zebrafish gene *cloche* acts upstream of a *flk-1* homologue to regulate endothelial cell differentiation. *Development* **124**, 381–389 (1997).
- Lyons, M. S., Bell, B., Stainier, D. & Peters, K. G. Isolation of the zebrafish homologues for the *tie-1* and *tie-2* endothelium-specific receptor tyrosine kinases. *Dev. Dyn.* **212**, 133–140 (1998).
- Lawson, N. D. & Weinstein, B. M. Arteries and veins: making a difference with zebrafish. *Nature Rev. Genet.* **3**, 674–682 (2002).
- Zhong, T. P., Rosenberg, M., Mohideen, M.-A. P. N. K., Weinstein, B. & Fishman, M. C. *gridlock*, an HLH gene required for assembly of the aorta in zebrafish. *Science* **287**, 1820–1824 (2000).
- Childs, S., Chen, J.-N., Garrity, D. M. & Fishman, M. C. Patterning of angiogenesis in the zebrafish embryo. *Development* **129**, 973–982 (2002).
- Sehnert, A. J. H. A., Weinstein, B. M., Walker, C., Fishman, M. & Stainier, D. Y. Cardiac troponin T is essential in sarcomere assembly and cardiac contractility. *Nature Genet.* **31**, 106–110 (2002).
- Isogai, S., Lawson, N. D., Torrealday, S., Horiguchi, M. & Weinstein, B. M. Angiogenic network

- formation in the developing vertebrate trunk. *Development* **130**, 5281–5290 (2003).
- Sumoy, L., Keasey, J. B., Dittman, T. D. & Kimelman, D. A role for notochord in axial vascular development revealed by analysis of phenotype and the expression of VEGF-2 in zebrafish *flh* and *ntl* mutant embryos. *Mech. Dev.* **63**, 15–27 (1997).
- Vokes, S. A. & Krieg, P. A. Endoderm is required for vascular endothelial tube formation, but not for angioblast specification. *Dev. Suppl.* **129**, 775–785 (2002).
- Odenthal, J. & Nusslein-Volhard, C. fork head domain genes in zebrafish. *Dev. Genes Evol.* **208**, 245–258 (1998).
- Schulte-Merker, S., Ho, R., Herrmann, B. & Nusslein-Volhard, C. The protein product of the zebrafish homologue of the mouse T gene is expressed in nuclei of the germ ring and the notochord of the early embryo. *Development* **116**, 1021–1032 (1992).
- Strahle, U., Blader, P., Henrique, D. & Ingham, P. Axial, a zebrafish gene expressed along the developing body axis, shows altered expression in cyclops mutant embryos. *Genes Dev.* **7**, 1436–1446 (1993).
- Parker, L. H., Zon, L. I. & Stainier, D. Y. Vascular and blood gene expression. *Methods Cell Biol.* **59**, 313–336 (1999).
- Fong, G. H., Zhang, L., Bryce, D. M. & Peng, J. Increased hemangioblast commitment, not vascular disorganization, is the primary defect in *flt-1* knock-out mice. *Development* **126**, 3015–3025 (1999).
- Ye, W., Shimamura, K., Rubenstein, J. L., Hynes, M. A. & Rosenthal, A. FGF and Shh signals control dopaminergic and serotonergic cell fate in the anterior neural plate. *Cell* **93**, 755–766 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. C. Fishman, B. Weinstein and N. Lawson for fish strains, plasmids and helpful discussions; N. Ferrara and H. Gerber for advice and for reviewing the manuscript; J. Lee for the zebrafish cDNA library; L. Rangell for electron microscopy; S. Greenwood for general lab assistance; R. Vanden, D. Yansura, R. Corpuz and H. Kim for recombinant EGFL7 proteins; A. Chuntharapai and C. Reed for monoclonal antibodies; and W. Wood, H. Clark and J. Tang for bioinformatics assistance. S.J. is supported by the American Heart Association. D.B. is a Human Frontier Science Program Organization fellow.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to W.Y. (loni@gene.com).

Costimulatory signals mediated by the ITAM motif cooperate with RANKL for bone homeostasis

Takako Koga^{1,2*}, Masanori Inui^{3*}, Kazuya Inoue³, Sunhwa Kim¹, Ayako Suematsu^{1,2}, Eiji Kobayashi³, Toshio Iwata³, Hiroshi Ohnishi⁴, Takashi Matozaki⁴, Tatsuhiko Kodama⁵, Tadatsugu Taniguchi¹, Hiroshi Takayanagi^{1,2} & Toshiyuki Takai³

¹Department of Immunology, Graduate School of Medicine and Faculty of Medicine, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033, Japan

²Department of Cellular Physiological Chemistry, Graduate School, Tokyo Medical and Dental University, and COE Program for Frontier Research on Molecular Destruction and Reconstruction of Tooth and Bone, Yushima 1-5-45, Bunkyo-ku, Tokyo 113-8549, Japan and PRESTO, Japan Science and Technology Agency (JST), Honcho 4-1-8, Kawaguchi, Saitama 332-0012, Japan

³Department of Experimental Immunology, Institute of Development, Aging and Cancer, Tohoku University, Seiryo 4-1, Aoba-ku, Sendai 980-8575, Japan and CREST, JST, Honcho 4-1-8, Kawaguchi, Saitama 332-0012, Japan

⁴Biosignal Research Center, Institute for Molecular and Cellular Regulation, Gunma University, Showa-machi 3-39-15, Maebashi 371-8512, Japan

⁵Department of Molecular Biology and Medicine, Research Center for Advanced Science and Technology, University of Tokyo, Komaba 4-6-1, Meguro-ku, Tokyo 153-8904, Japan

* These authors contributed equally to this work

Costimulatory signals are required for activation of immune cells¹, but it is not known whether they contribute to other biological systems. The development and homeostasis of the skeletal system depend on the balance between bone formation and resorption^{2,3}. Receptor activator of NF-κB ligand (RANKL) regulates the differentiation of bone-resorbing cells, osteoclasts, in the presence of macrophage-colony stimulating factor

(M-CSF)^{4,5}. But it remains unclear how RANKL activates the calcium signals that lead to induction of nuclear factor of activated T cells c1, a key transcription factor for osteoclastogenesis⁶. Here we show that mice lacking immunoreceptor tyrosine-based activation motif (ITAM)⁷-harbouring adaptors^{8–10}, Fc receptor common γ subunit (FcR γ) and DNAX-activating protein (DAP)12, exhibit severe osteopetrosis owing to impaired osteoclast differentiation. In osteoclast precursor cells, FcR γ and DAP12 associate with multiple immunoreceptors^{11–15} and activate calcium signalling through phospholipase C γ . Thus, ITAM-dependent costimulatory signals activated by multiple immunoreceptors are essential for the maintenance of bone homeostasis. These results reveal that RANKL and M-CSF are not sufficient to activate the signals required for osteoclastogenesis.

Osteoclast differentiation induced by RANKL and M-CSF is severely blocked in the bone marrow monocyte/macrophage lineage cells (BMMs) derived from mice lacking DAP12 (ref. 16), also known as KARAP or TYROBP^{9,10}. DAP12 is a membrane adaptor molecule that contains an ITAM motif, which activates calcium signalling in immune cells. Despite this *in vitro* blockage, DAP12-deficient (*DAP12*^{-/-}) mice exhibit only mild osteopetrosis and contain a normal number of osteoclasts. This suggests that the DAP12-mediated signal plays a crucial role in the RANKL/M-CSF-induced osteoclast formation system but that another molecule(s) can rescue the DAP12 deficiency *in vivo*¹⁶.

To investigate the molecular basis of this discrepancy, we stimulated BMMs with RANKL and M-CSF after strictly depleting the stromal/osteoblastic cells¹⁷. In this osteoblast-free system, osteoclast differentiation is completely abrogated in *DAP12*^{-/-} BMMs (Fig. 1a); we observed no formation of multinucleated cells (MNCs) positive for the marker enzyme of osteoclasts, tartrate-resistant acid phosphatase (TRAP). This differentiation block is efficiently rescued by retroviral expression of DAP12, but not by DAP12Y65F, a DAP12 mutant that does not transmit the ITAM signal¹⁸ (Fig. 1b), indicating that the DAP12-mediated ITAM signal is required for osteoclastogenesis in this system. Interestingly, *DAP12*^{-/-} BMMs undergo osteoclast differentiation in co-culture with osteoblasts¹⁹ (Fig. 1c), showing that osteoblasts can stimulate the signal that compensates for the loss of the DAP12-mediated ITAM signal. This compensatory mechanism may explain the normal osteoclast differentiation in *DAP12*^{-/-} mice *in vivo* (see Supplementary Fig. 1a, b).

The osteoclast-associated receptor (OSCAR) is an activating-type immunoglobulin-like receptor induced in RANKL-stimulated BMMs¹¹. Expression of a putative OSCAR ligand in osteoblasts¹¹ led us to investigate OSCAR as a candidate receptor that compensates for the loss of DAP12-mediated signalling. To explore the adaptor molecules with which OSCAR associates, we constructed a chimaeric receptor composed of the extracellular portion of type IIB FcR for IgG (Fc γ RIIB) and the transmembrane and cytoplasmic portion of OSCAR (RIIB-OSCAR). Flow-cytometric analysis revealed that RIIB-OSCAR expression was significantly enhanced only when it was cotransfected with FcR γ (Fig. 1d). Thus, RIIB-OSCAR preferentially associates with FcR γ , another adaptor protein containing an ITAM motif, but not with DAP12 or DAP10. Immunoprecipitation analysis also verified the exclusive association of RIIB-OSCAR with FcR γ (Supplementary Fig. 1c).

The results prompted us to investigate the bone phenotype of mice deficient in FcR γ (*FcR γ* ^{-/-} mice)²⁰, but there was no significant difference in the osteoclast number and the trabecular bone volume between wild-type and *FcR γ* ^{-/-} mice (Fig. 1e and Supplementary Fig. 1d). In addition, there was little, if any, difference between wild-type and *FcR γ* ^{-/-} mice in the differentiation of osteoclasts in the RANKL/M-CSF system (Supplementary Fig. 1e) and the co-culture system (data not shown). Although OSCAR-Fc,

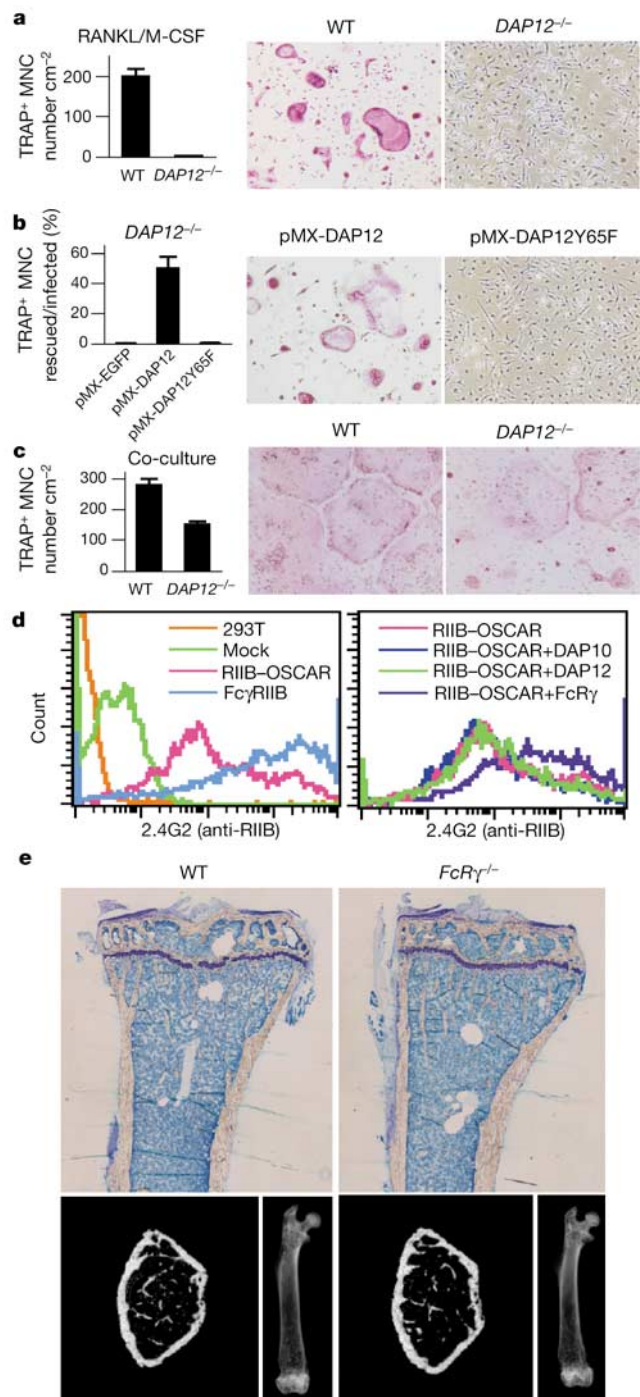


Figure 1 Impaired osteoclastogenesis in the absence of DAP12 and the compensatory mechanism by osteoblasts. **a**, Complete lack of osteoclastogenesis in *DAP12*^{-/-} BMMs stimulated with RANKL/M-CSF. **b**, Rescue of osteoclastogenesis by retrovirus-mediated expression of DAP12, but not DAP12Y65F, in *DAP12*^{-/-} BMMs. These multinucleated cells have a bone-resorbing activity (not shown). **c**, Osteoclast formation from *DAP12*^{-/-} BMMs in co-culture with osteoblasts. Bone-resorbing activity was examined on dentine slices (not shown). **d**, Flow-cytometric analysis of cell-surface expression of RIIB-OSCAR or Fc γ RIIB after overexpression of the chimaeric receptor of RIIB-OSCAR or Fc γ RIIB (left). Weak expression of Fc γ RIIB is detected by transfection of RIIB-OSCAR alone (pink line, left and right panels), but cotransfection with FcR γ exclusively increased this expression (violet line, right). **e**, Histology of tibia (upper, toluidine blue staining) and microradiographic analysis of femur (lower, microcomputed tomography, left; microradiograph, right) of *FcR γ* ^{-/-} mice. There was no significant abnormality.

a fusion protein of OSCAR ectodomain and Fc portion of IgG, suppresses osteoclastogenesis *in vitro*¹¹, our results suggest that FcR γ -mediated signals can be compensated for by other signals *in vivo*.

Considering the possibility that DAP12 and FcR γ functionally compensate for each other²¹, we generated mice lacking both molecules (DAP12^{-/-} FcR γ ^{-/-} mice). These mice exhibit severe osteopetrosis, with bone marrow filled with unresorbed bone (Fig. 2a). DAP12^{-/-} mice show mild osteopetrosis due to impaired osteoclast activity¹⁶, but the osteopetrosis in DAP12^{-/-} FcR γ ^{-/-} mice is much more severe, as seen in microradiographs (Fig. 2a) and trabecular bone volume (Fig. 2b). Importantly, we observed few osteoclasts in DAP12^{-/-} FcR γ ^{-/-} mice, indicating that the osteopetrosis is caused by defective differentiation rather than by defective activity of osteoclasts (Fig. 2c, e). Bone-morphometric analysis revealed that osteoblastic bone formation also decreases (Fig. 2d and Supplementary Fig. 2a). Thus, the ITAM-harboring adaptors FcR γ and DAP12 are essential for osteoclast differentiation *in vivo*. Despite the severe osteopetrosis in DAP12^{-/-} FcR γ ^{-/-} mice, these mice have no defect in tooth eruption (data not shown). In addition, we observed a very small number of TRAP⁺ MNCs in limited areas just below the epiphyseal plate (Fig. 2e and Supplementary Fig. 2b), suggesting that another adaptor molecule(s) may compensate for the function under specific conditions. In the culture system, DAP12^{-/-} FcR γ ^{-/-} osteoclast precursor cells cannot undergo osteoclast differentiation in response to RANKL and M-CSF (Fig. 3a). Retroviral transfer of DAP12, but not DAP12Y65F or FcR γ , into DAP12^{-/-} FcR γ ^{-/-} cells efficiently rescued osteoclast differentiation induced by RANKL and M-CSF. In addition, DAP12^{-/-} FcR γ ^{-/-} precursor cells barely differentiate into osteoclasts even when co-cultured with osteoblasts (Fig. 3b). In the co-culture system, retroviral transfer of FcR γ , but not ITAM-deficient FcR γ Y65F, also rescued osteoclast differentiation, albeit partially. These results suggest that the ITAM signal mediated through FcR γ and DAP12 is indispensable for RANKL-induced osteoclastogenesis and that each adaptor-mediated signal is differentially regulated.

FcR γ or DAP12 associates with several specific immunoreceptors for cell activation in myeloid lineage cells^{13,21,22}. To identify receptors that associate with FcR γ and DAP12 in osteoclast lineage cells, we screened the expression profiles of messenger RNAs for known candidate receptors using GeneChip analysis. We found that a series of receptors and ITAM-associated molecules, as well as FcR γ and DAP12, are expressed in the osteoclast lineage (Supplementary Fig. 3a). These putative FcR γ - and DAP12-associating receptors¹¹⁻¹⁵ were detected on the surface of the osteoclast lineage using cell-surface labelling with biotin (Supplementary Fig. 3b, c). Among these, immunoprecipitation experiments confirmed that paired immunoglobulin-like receptor (PIR)-A, Fc γ RIII and OSCAR each pair with FcR γ and that triggering receptor expressed by myeloid cells (TREM)-2 and signal-regulatory protein (SIRP) β 1, but not NKG2D, pair with DAP12 in the osteoclast lineage (Fig. 3c).

To test whether these receptor-mediated signals promote osteoclast differentiation by associating with FcR γ or DAP12, we stimulated BMMs with plate-bound monoclonal antibodies against OSCAR, PIR, TREM-2 and SIRP β 1. Triggering of either receptor by crosslinking with an antibody accelerated RANKL-induced osteoclast differentiation, indicating that these receptors activate osteoclastogenesis, although the stimulatory effect is more obvious at a low concentration of RANKL (Fig. 3d and Supplementary Fig. 3d). In the absence of RANKL, the stimulation of these receptors alone could not induce osteoclast differentiation (data not shown), suggesting that these receptor-mediated signals act cooperatively with RANKL but cannot substitute for the signal.

A stimulating effect by anti-OSCAR and anti-PIR antibodies was not observed in FcR γ ^{-/-} BMMs, whereas anti-TREM-2 and anti-

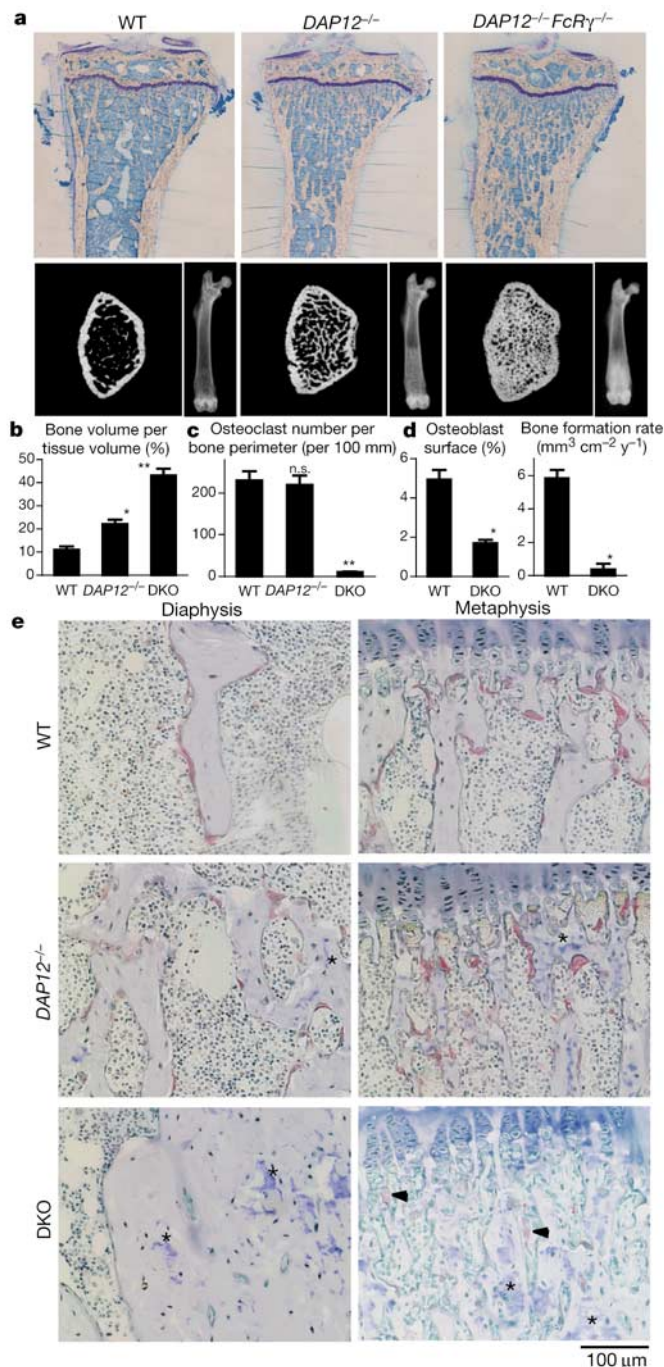


Figure 2 Severe osteopetrosis in DAP12^{-/-} FcR γ ^{-/-} (DKO) mice due to impaired osteoclast differentiation. **a**, Histology of tibia and microradiographic analysis of femur of DAP12^{-/-} and DAP12^{-/-} FcR γ ^{-/-} mice (12 weeks of age). Bone-marrow cavity is absent in DAP12^{-/-} FcR γ ^{-/-} mice. Higher radiopacity shows that DAP12^{-/-} FcR γ ^{-/-} mice have a much more severe osteopetrotic phenotype than DAP12^{-/-} mice. **b**, Increased trabecular bone volume in DAP12^{-/-} FcR γ ^{-/-} mice. **c**, The number of osteoclasts is not altered in DAP12^{-/-} mice, but is markedly decreased in DAP12^{-/-} FcR γ ^{-/-} mice. **d**, Osteoblastic parameters in the bone-morphometric analysis of the tibia of wild-type and DAP12^{-/-} FcR γ ^{-/-} mice (12 weeks of age). **e**, Histology of the tibia of DAP12^{-/-} and DAP12^{-/-} FcR γ ^{-/-} mice (TRAP and toluidine blue staining). Bone-marrow cavity is filled with unresorbed bone in the diaphysis of DAP12^{-/-} FcR γ ^{-/-} mice. No osteoclasts are observed in the diaphysis of DAP12^{-/-} FcR γ ^{-/-} mice, whereas the osteoclast number is normal in DAP12^{-/-} mice. Typical sites of cartilage remnant are indicated by asterisks. In the metaphyseal area, a small number of osteoclasts are observed below epiphyseal plates in DAP12^{-/-} FcR γ ^{-/-} mice (arrowheads).

SIRP β 1 antibodies stimulated *FcR γ ^{-/-}* BMMs, the same as they did with wild-type cells (Fig. 3e). This indicates that *FcR γ* is required for the stimulatory function of OSCAR and PIR-A. Consistent with the observation that osteoblasts rescue *DAP12* deficiency through *FcR γ* -associated receptors, anti-OSCAR and anti-PIR antibodies could rescue the osteoclastogenesis from *DAP12*^{-/-} BMMs but not from *DAP12*^{-/-} *FcR γ ^{-/-}* cells (Fig. 3f). Stimulation of *DAP12*-associating receptors such as TREM-2 and SIRP β 1 did not rescue *DAP12*^{-/-} BMMs. This indicates that *DAP12* is required for the stimulatory effect through TREM-2 and SIRP β 1 (Fig. 3f).

How does the ITAM signal contribute to RANKL-induced signalling events? To address this question, we performed a genome-wide screening of mRNA expression in RANKL-stimulated osteoclast precursor cells derived from wild-type and *DAP12*^{-/-} *FcR γ ^{-/-}* mice. Among the transcription factors and effector molecules involved in RANKL signalling⁶, the induction of nuclear factor of activated T cells c1 (NFATc1) is most strongly suppressed in *DAP12*^{-/-} *FcR γ ^{-/-}* cells (Fig. 4a). We analysed the protein expression of essential molecules for osteoclastogenesis^{2,4,6} including NFATc1, c-Fos and TRAF6, and revealed that NFATc1 expression in *DAP12*^{-/-} *FcR γ ^{-/-}* cells stimulated with RANKL is barely detectable, but the expression of c-Fos or TRAF6 is still observed under the same conditions (Supplementary Fig. 4a). To

confirm the crucial role of NFATc1 as a downstream target, we examined whether ectopic expression of NFATc1, c-Fos or TRAF6 by retrovirus-mediated gene transfer rescues the differentiation block of osteoclasts in *DAP12*^{-/-} *FcR γ ^{-/-}* cells. Ectopic expression of NFATc1, but not c-Fos or TRAF6, resulted in efficient osteoclast formation even in *DAP12*^{-/-} *FcR γ ^{-/-}* cells (Fig. 4b).

RANKL-induced calcium signalling is essential for autoamplification of the *NFATc1* gene during osteoclastogenesis⁶, whereas immunoglobulin-like receptors activate phospholipase C γ (PLC γ) and calcium signalling through ITAM in immune cells^{13,21}. We therefore examined calcium signalling in RANKL-stimulated osteoclast precursor cells derived from *DAP12*^{-/-} *FcR γ ^{-/-}* mice. As shown in Fig. 4c, calcium oscillation induced by RANKL is not significantly observed in *DAP12*^{-/-} *FcR γ ^{-/-}* cells, which suggests that the inhibition of this calcium signalling explains the impaired induction of NFATc1. Furthermore, stimulation with the anti-PIR antibody rescued the defect in calcium signalling and NFATc1 expression in *DAP12*^{-/-} cells, suggesting that the stimulation of an immunoreceptor is required for RANKL-induced activation of the calcium signal leading to NFATc1 induction and osteoclastogenesis (Fig. 4c and Supplementary Fig. 4b). These findings suggest that *FcR γ* and *DAP12* are required for the induction of NFATc1, the crucial step in the RANKL-induced osteoclast

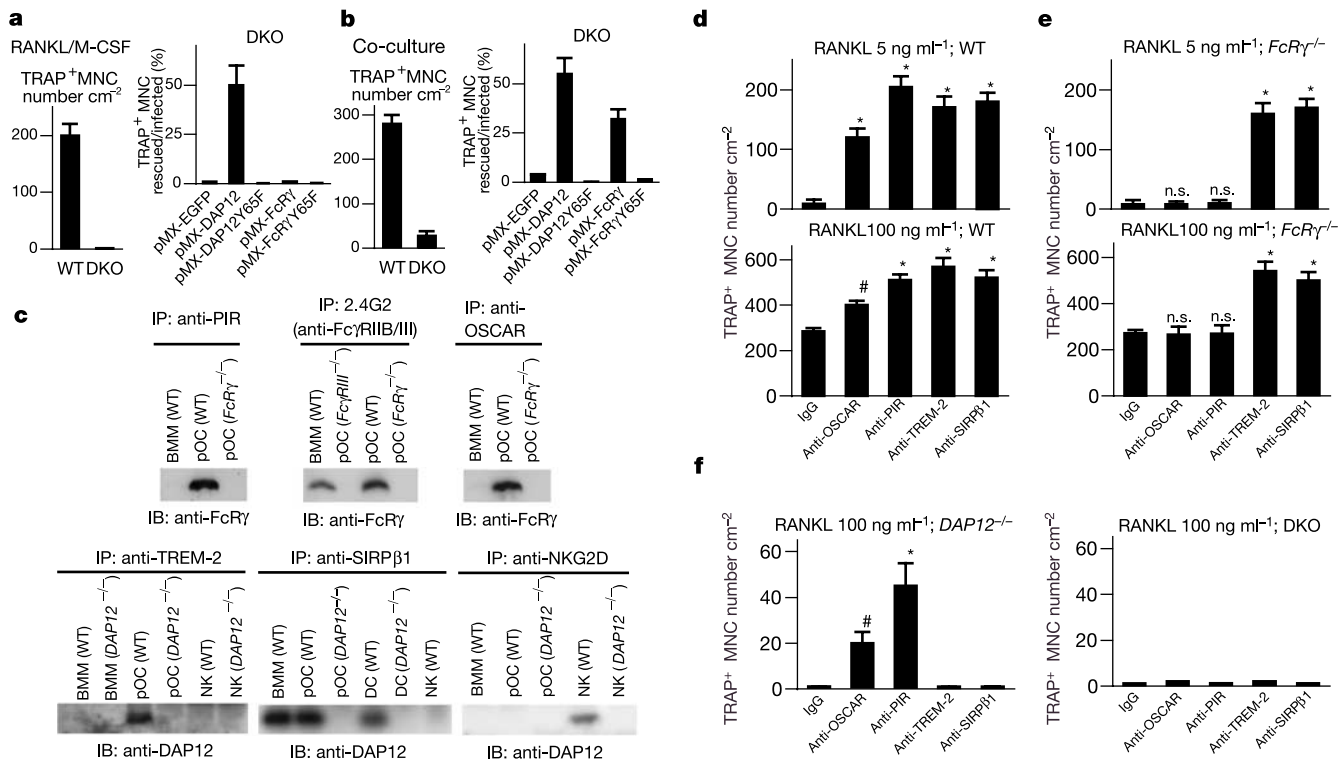


Figure 3 Distinct roles of *FcR γ* - and *DAP12*-associating immunoreceptors in the regulation of osteoclastogenesis. **a**, Complete block of *in vitro* osteoclastogenesis in *DAP12*^{-/-} *FcR γ ^{-/-}* (DKO) cells stimulated with RANKL and M-CSF. This was rescued by pMX-DAP12, but not pMX-DAP12Y65F or pMX-*FcR γ* . **b**, Severe impairment of osteoclastogenesis in *DAP12*^{-/-} *FcR γ ^{-/-}* cells even in co-culture with osteoblasts. Osteoclastogenesis was rescued by pMX-DAP12 or pMX-*FcR γ* in an ITAM-dependent manner. **c**, Association of *FcR γ* with PIR-A, *FcγRIIb/III* and OSCAR (upper). pOC, osteoclast precursor cells stimulated with RANKL. For the comments on the association of *FcR γ* with *FcγRIIb/III*, see Supplementary Fig. 3b. **c**, Association of *DAP12* with TREM-2 and SIRP β 1, but not with NKG2D, in pOC (lower). For the positive control, *DAP12* associates with SIRP β 1 and NKG2D in dendritic cells (DC) and NK cells (NK), respectively. **d**, Treatment of

BMMs with plate-bound monoclonal antibodies against OSCAR, PIR, TREM-2 or SIRP β 1 promotes osteoclastogenesis stimulated by RANKL and M-CSF. The stimulatory effect by antibody-mediated crosslinking is more obvious at a low concentration of RANKL (5 ng ml⁻¹) than at a high concentration (100 ng ml⁻¹). **e**, Effect of plate-bound antibodies on osteoclastogenesis from *FcR γ ^{-/-}* BMMs. Anti-OSCAR and anti-PIR antibodies have no effect on *FcR γ ^{-/-}* BMMs. **f**, Effect of plate-bound antibodies on osteoclastogenesis from *DAP12*^{-/-} BMMs or *DAP12*^{-/-} *FcR γ ^{-/-}* cells stimulated with RANKL/M-CSF. Anti-OSCAR and anti-PIR antibodies rescued osteoclastogenesis from *DAP12*^{-/-} BMMs, but such an effect was not observed in anti-TREM-2 or anti-SIRP β 1 antibodies. No stimulatory effect was observed on osteoclastogenesis from *DAP12*^{-/-} *FcR γ ^{-/-}* cells.

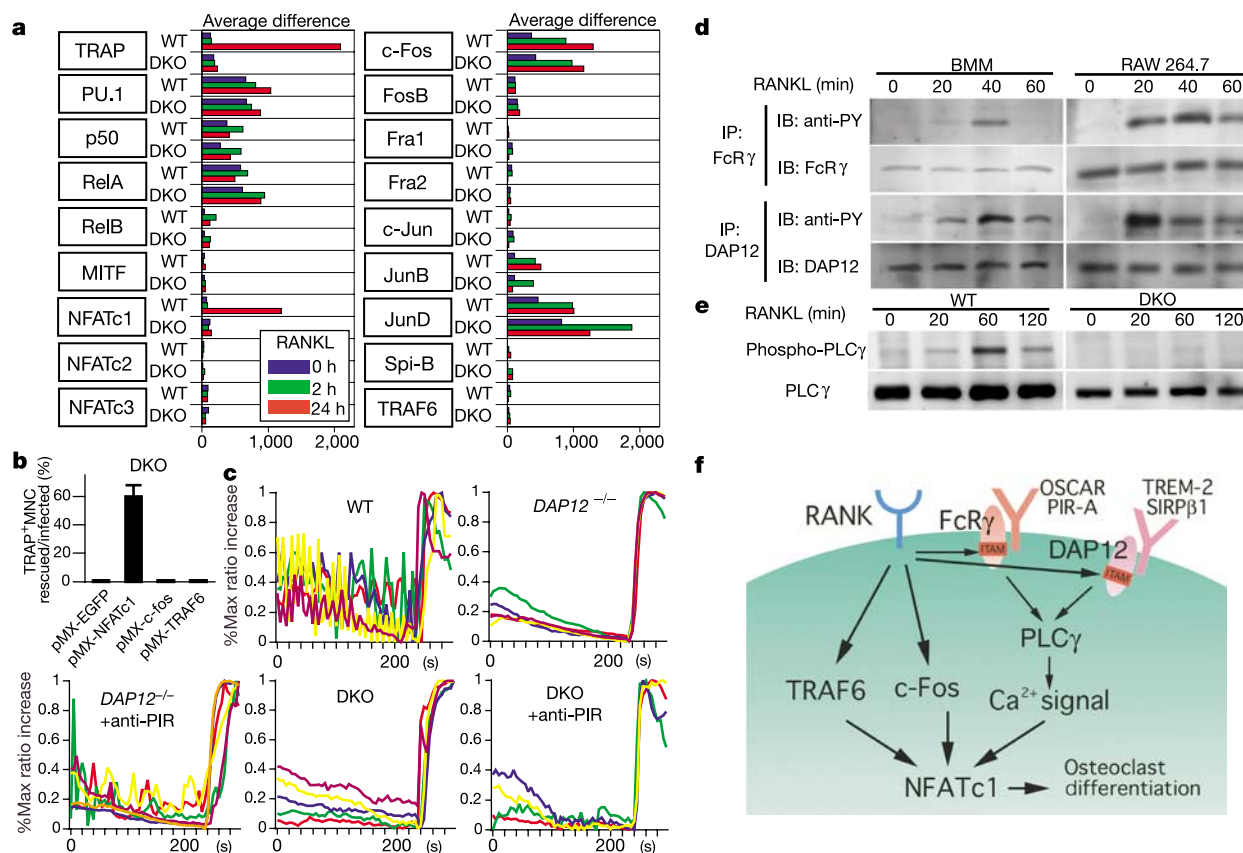


Figure 4 ITAM-harboring adaptors are essential for RANKL induction of calcium signalling and NFATc1. **a**, GeneChip analysis of mRNA expression of transcription factors and effector molecules involved in RANKL signalling. NFATc1 induction is most strongly impaired in $DAP12^{-/-}$ $FcR\gamma^{-/-}$ (DKO) cells. **b**, Introduction of NFATc1 by pMX retrovirus vector (pMX-NFATc1) into the $DAP12^{-/-}$ $FcR\gamma^{-/-}$ precursor cells efficiently rescued the osteoclastogenesis (with bone-resorbing activity; not shown) in the presence of RANKL and M-CSF. **c**, Calcium signalling in osteoclast precursor cells stimulated with RANKL and M-CSF for 24 h. Calcium oscillation was not observed in $DAP12^{-/-}$ $FcR\gamma^{-/-}$ or $DAP12^{-/-}$ cells. Stimulation with plate-bound anti-PIR antibody rescued the calcium signalling in $DAP12^{-/-}$ cells, but not in $DAP12^{-/-}$ $FcR\gamma^{-/-}$ cells. **d**,

Phosphorylation of FcR γ and DAP12 by RANKL. Phosphorylation of FcR γ and DAP12 induced by RANKL in BMMs as well as in RAW 264.7 cells. PY, phosphotyrosine. **e**, Impaired phosphorylation of PLC γ 1 by RANKL in $DAP12^{-/-}$ $FcR\gamma^{-/-}$ cells. **f**, A schematic model of ITAM-mediated costimulatory signal in RANKL-stimulated induction of osteoclast differentiation. Phosphorylation of ITAM stimulated by immunoreceptors and RANKL–RANK interaction results in the recruitment of Syk family kinases, leading to the activation of PLC γ and calcium signalling, which is critical for NFATc1 induction. NFATc1 induction is also dependent on c-Fos and TRAF6, both of which are activated by RANKL. RANKL may also contribute to efficient ITAM signalling through the induction of immunoreceptors or their putative ligands.

differentiation programme, through activation of calcium signalling.

To gain further insight into the mechanism linking RANKL and ITAM-mediated calcium signalling, we analysed the phosphorylation events induced by RANKL. Our analysis revealed that RANKL induces the phosphorylation of both FcR γ and DAP12 in BMMs as well as in RAW 264.7 cells (Fig. 4d). We further demonstrated that RANKL-induced phosphorylation of PLC γ is impaired in $DAP12^{-/-}$ $FcR\gamma^{-/-}$ cells (Fig. 4e), but RANKL-induced phosphorylation of p38, JNK and I κ B is not affected (Supplementary Fig. 4c). This suggests that FcR γ and DAP12 are specifically involved in the PLC γ –calcium pathway. In addition, piceatannol, the inhibitor of Syk kinase, which is recruited to ITAM in immune cells, has an inhibitory effect on osteoclastogenesis (Supplementary Fig. 4d).

Our results show that the ITAM-harboring adaptors FcR γ and DAP12 deliver essential signals, in concert with RANKL-induced signalling cascades, for terminal differentiation of osteoclasts. Our study establishes the importance of the ITAM-mediated costimulatory signal in RANKL-induced osteoclast differentiation (Fig. 4f), which is analogous to the role of costimulatory signals in the immune system^{13,21}. Although most of the ligands for the immuno-

receptors remain to be identified, putative ligands are provided by distinct cell types: osteoclast precursor cells themselves stimulate exclusively DAP12-associated receptors, but osteoblasts also stimulate FcR γ -associated receptors. The defective osteoblast function in $DAP12^{-/-}$ $FcR\gamma^{-/-}$ mice suggests that the osteoclast–osteoblast communication through these receptors may couple bone resorption to formation. It remains to be elucidated whether inhibitory receptors such as Fc γ RIIB and PIR-B (Supplementary Fig. 3a) contribute to counterbalancing the ITAM signal in osteoclast differentiation. It is notable that osteoclastogenesis is enhanced in mice lacking phosphatases such as SHP-1 or SHIP-1, which counterbalance the ITAM signal in the immune system^{23,24}. The mutations of the ITAM-harboring adaptors and their associating receptors are related to pathological conditions in the skeletal system^{25–27}, suggesting that the regulation of costimulatory signals may provide a novel strategy for the treatment of skeletal disorders. □

Methods

Mice and analysis of bone phenotype

$DAP12^{-/-}$ $FcR\gamma^{-/-}$ mice were generated by crossing between $DAP12^{+/-}$ $FcR\gamma^{+/-}$ males and females obtained by mating of $DAP12^{-/-}$ mice in the 129/SvJ and C57BL/6 (B6)

hybrid background¹⁶ with *FcRγ*^{-/-} mice in B6 background²⁰. *DAP12*^{-/-} *FcRγ*^{-/-} mice grow normally and are fertile, without gross abnormalities. *FcγRIII*^{-/-} mice have been described previously²⁸. All experiments were performed with appropriate littermate controls. Histological, histomorphometric and microradiographic examinations were performed using essentially the same method as described previously²⁹. Statistical analysis was performed using the Student's *t*-test ($\#P < 0.05$, $*P < 0.01$, $**P < 0.001$). All mice were born and kept under pathogen-free conditions.

Flow cytometry and immunoprecipitation

Complementary DNAs for *FcRγ*, *DAP12*, *DAP10* and *RIIB*-OSCAR chimaera coding for the extracellular domain of *FcγRIIB* and the transmembrane and cytoplasmic portion of OSCAR were inserted into the pIRES-puro vector (BD Biosciences). 293T cells were transfected transiently with these vectors, harvested after 48 h, stained with PE-conjugated anti-*FcγRIIB/III* antibody (2.4G2) and monitored by flow cytometry. For immunoprecipitation, cells were lysed with digitonin buffer (1% digitonin, 13.6 mM triethanolamine, 150 mM NaCl, 1 mM EDTA, 10 mM iodoacetamide, protease inhibitors, pH 7.8). Cell lysates were incubated with 2.4G2 (ATCC), anti-OSCAR, anti-TREM-2 (6E9) (prepared by T. Takai), anti-PIR (6C1, a gift from H. Kubagawa), anti-NKG2D (Santa Cruz) and anti-SIRPβ1 (prepared by T. M.) antibodies and Protein A Sepharose. The specificity of newly prepared antibodies was determined using flow cytometry, and their agonistic effects on ITAM-mediated signalling events were confirmed. Immunoprecipitates were separated by SDS-polyacrylamide gel electrophoresis (PAGE) and immunoblotted with anti-*FcRγ* or -*DAP12* antibodies (prepared by T. Takai).

In vitro osteoclastogenesis

We have described our method of *in vitro* osteoclastogenesis previously^{17,29}. Briefly, after depleting adherent cells, non-adherent bone marrow cells were cultured in α -MEM (Gibco BRL) with 10% FBS (Sigma) containing 10 ng ml⁻¹ M-CSF (Genzyme). After two days, adherent cells were used as BMMs. Monocyte/macrophage progenitor cells of *DAP12*^{-/-} *FcRγ*^{-/-} mice were derived from the spleen and similarly cultured with M-CSF for 2 days. These osteoclast precursor cells were further cultured in the presence of 100 ng ml⁻¹ soluble RANKL (Peprotech) and 10 ng ml⁻¹ M-CSF to generate osteoclasts (RANKL/M-CSF system). RANKL/M-CSF were used at these concentrations throughout the paper unless otherwise described. To examine the *in vitro* osteoclastogenesis from *DAP12*^{-/-} BMMs, the contamination of stromal/osteoblastic cells should be strictly avoided¹⁷. These culture conditions are different from those previously described¹⁶, in which osteoclast formation was not completely blocked. The co-culture of osteoclast precursor cells and osteoblasts derived from calvarial cells was performed in the presence of 10⁻⁸ M 1, 25 (OH)₂ vitamin D₃ and 10⁻⁶ M dexamethasone in the absence of recombinant RANKL and M-CSF as described previously¹⁹. Three to five days later, TRAP⁺ multinucleated (more than three nuclei) cells were counted. All data are expressed as mean $\pm$ s.e.m. ($n = 6$). TRAP⁺ MNCs were characterized by examining the bone-resorbing activity on dentine slices as described previously⁶. For crosslinking experiments, culture plates were preincubated with PBS containing 5 μ g ml⁻¹ antibody at 4 °C for 24 h before seeding precursor cells.

Retroviral gene transduction

Retroviral vectors pMX-DAP12, pMX-DAP12Y65F, pMX-*FcRγ* and pMX- *FcRγ*Y65F were constructed by inserting *DAP12* or *FcRγ* cDNA, and their mutant cDNAs generated by PCR-directed mutagenesis into pMX-IRES-EGFP vector¹⁷. Other vectors such as pMX-NFATc1, pMX-*c-fos* and pMX-TRAF6 have been described previously, and packaging was performed as described elsewhere^{6,29}. Two days after inoculation, BMMs were cultured with RANKL and M-CSF. After four days, osteoclastogenesis was evaluated by TRAP staining. The rescuing effect was normalized by measuring infection efficiency assessed by GFP expression as previously described^{6,29}.

GeneChip analysis

RNA extraction was previously described⁶. Total RNA (15 μ g) was used for cDNA synthesis by reverse transcription followed by synthesis of biotinated cRNA through *in vitro* transcription. After cRNA fragmentation, hybridization with mouse U74Av2 or A430 GeneChip (Affymetrix) was performed and analysed according to the manufacturer's protocol. GeneChip analysis was repeated several times and yielded similar results; a representative set of data is shown.

Calcium measurement

Osteoclast precursor cells were incubated with RANKL in the presence of M-CSF for 24 h and subjected to calcium measurement as previously described⁶. RANKL does not induce a calcium spike immediately in the osteoclast precursor cells, although it does so in the mature osteoclasts^{6,30}.

Phosphorylation of *FcRγ*/DAP12 and PLCγ

Osteoclast precursor cells or RAW 264.7 cells were stimulated by 100 ng ml⁻¹ soluble RANKL after 6 h of serum starvation. After various time periods, cell extracts were harvested from the cells using TNE buffer containing 10 mM Tris-HCl (pH 7.8), 150 mM NaCl, 1 mM EDTA, 1% NP-40, 2 mM Na₃VO₄, 10 mM NaF and 10 μ g ml⁻¹ aprotinin. Cell extracts were incubated with 1 μ g of anti-DAP12 or anti-*FcRγ* antibodies for 1 h at 4 °C. Immune complexes were recovered with Protein A Sepharose, subjected to SDS-PAGE and blotted with anti-phosphotyrosine antibody (4G10, Upstate) or the indicated antibodies. Activation of PLCγ was detected using anti-PLCγ1 and anti-phospho-PLCγ1 antibodies (Santa Cruz).

Received 2 December 2003; accepted 23 February 2004; doi:10.1038/nature02444.

- Lenschow, D. J., Walunas, T. L. & Bluestone, J. A. CD28/B7 system of T cell costimulation. *Annu. Rev. Immunol.* **14**, 233–258 (1996).
- Karsenty, G. & Wagner, E. F. Reaching a genetic and molecular understanding of skeletal development. *Dev. Cell.* **2**, 389–406 (2002).
- Rodan, G. A. & Martin, T. J. Therapeutic approaches to bone diseases. *Science* **289**, 1508–1514 (2000).
- Teitelbaum, S. L. & Ross, F. P. Genetic regulation of osteoclast development and function. *Nature Rev. Genet.* **4**, 638–649 (2003).
- Boyle, W. J., Simonet, W. S. & Lacey, D. L. Osteoclast differentiation and activation. *Nature* **423**, 337–342 (2003).
- Takayanagi, H. *et al.* Induction and activation of the transcription factor NFATc1 (NFAT2) integrate RANKL signaling in terminal differentiation of osteoclasts. *Dev. Cell.* **3**, 889–901 (2002).
- Reth, M. Antigen receptor tail clue. *Nature* **338**, 383–384 (1989).
- Perez-Montfort, R., Kinet, J. P. & Metzger, H. A previously unrecognized subunit of the receptor for immunoglobulin E. *Biochemistry* **22**, 5722–5728 (1983).
- Olcese, L. *et al.* Human killer cell activatory receptors for MHC class I molecules are included in a multimeric complex expressed by natural killer cells. *J. Immunol.* **158**, 5083–5086 (1997).
- Lanier, L. L., Corliss, B. C., Wu, J., Leong, C. & Phillips, J. H. Immunoreceptor DAP12 bearing a tyrosine-based activation motif is involved in activating NK cells. *Nature* **391**, 703–707 (1998).
- Kim, N., Takami, M., Rho, J., Josien, R. & Choi, Y. A novel member of the leukocyte receptor complex regulates osteoclast differentiation. *J. Exp. Med.* **195**, 201–209 (2002).
- Kubagawa, H., Burrows, P. D. & Cooper, M. D. A novel pair of immunoglobulin-like receptors expressed by B cells and myeloid cells. *Proc. Natl Acad. Sci. USA* **94**, 5261–5266 (1997).
- Colonna, M. TREMs in the immune system and beyond. *Nature Rev. Immunol.* **3**, 445–453 (2003).
- Dietrich, J., Cella, M., Seiffert, M., Bühring, H. J. & Colonna, M. Cutting edge: signal-regulatory protein β 1 is a DAP12-associated activating receptor expressed in myeloid cells. *J. Immunol.* **164**, 9–12 (2000).
- Tomasello, E. *et al.* Association of signal-regulatory proteins β with KARAP/DAP-12. *Eur. J. Immunol.* **30**, 2147–2156 (2000).
- Kaifu, T. *et al.* Osteopetrosis and thalamic hypomyelination with synaptic degeneration in DAP12-deficient mice. *J. Clin. Invest.* **111**, 323–332 (2003).
- Takayanagi, H. *et al.* T-cell-mediated regulation of osteoclastogenesis by signalling cross-talk between RANKL and IFN- γ . *Nature* **408**, 600–605 (2000).
- Tomasello, E. *et al.* Gene structure, expression pattern, and biological activity of mouse killer cell activating receptor-associated protein (KARAP)/DAP-12. *J. Biol. Chem.* **273**, 34115–34119 (1998).
- Takahashi, N. *et al.* Osteoblastic cells are involved in osteoclast formation. *Endocrinology* **123**, 2600–2602 (1988).
- Takai, T., Li, M., Sylvestre, D., Clynes, R. & Ravetch, J. V. *FcRγ* chain deletion results in pleiotropic effector cell defects. *Cell* **76**, 519–529 (1994).
- Takai, T. Roles of Fc receptors in autoimmunity. *Nature Rev. Immunol.* **2**, 580–592 (2002).
- Cerwenka, A. & Lanier, L. L. Natural killer cells, viruses and cancer. *Nature Rev. Immunol.* **1**, 41–49 (2001).
- Aoki, K. *et al.* The tyrosine phosphatase SHP-1 is a negative regulator of osteoclastogenesis and osteoclast resorbing activity: increased resorption and osteopenia in *me^o/me^o* mutant mice. *Bone* **25**, 261–267 (1999).
- Takeshita, S. *et al.* SHIP-deficient mice are severely osteoporotic due to increased numbers of hyper-resorptive osteoclasts. *Nature Med.* **8**, 943–949 (2002).
- Paloneva, J. *et al.* Mutations in two genes encoding different subunits of a receptor signaling complex result in an identical disease phenotype. *Am. J. Hum. Genet.* **71**, 656–662 (2002).
- Paloneva, J. *et al.* DAP12/TREM2 deficiency results in impaired osteoclast differentiation and osteoporotic features. *J. Exp. Med.* **198**, 669–675 (2003).
- Cella, M. *et al.* Impaired differentiation of osteoclasts in TREM-2-deficient individuals. *J. Exp. Med.* **198**, 645–651 (2003).
- Ujike, A. *et al.* Modulation of immunoglobulin (Ig)E-mediated systemic anaphylaxis by low-affinity Fc receptors for IgG. *J. Exp. Med.* **189**, 1573–1579 (1999).
- Takayanagi, H. *et al.* RANKL maintains bone homeostasis through c-Fos-dependent induction of interferon- β . *Nature* **416**, 744–749 (2002).
- Komarova, S. V., Pilkington, M. F., Weidema, A. E., Dixon, S. J. & Sims, S. M. RANK ligand-induced elevation of cytosolic Ca²⁺ accelerates nuclear translocation of nuclear factor κ B in osteoclasts. *J. Biol. Chem.* **278**, 8286–8293 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. V. Ravetch, H. Kubagawa and M. D. Cooper for providing materials, and M. Kaji, A. Sugahara, Y. Ito, K. Maya, A. Sato, A. Nakamura, M. Isobe, T. Yokochi, A. Izumi, T. Kohro, Y. Matsui, H. Murayama, K. Sato, M. Asagiri and I. Kawai for technical assistance and discussion. This work was supported in part by a grant for Advanced Research on Cancer from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, the CREST and PRESTO programs of Japan Science and Technology Agency (JST), grants for the 21st century COE program 'Frontier Research on Molecular Destruction and Reconstruction of Tooth and Bone' and 'Center for Innovative Therapeutic Development Towards the Conquest of Signal Transduction Diseases', Grants-in-Aid for Scientific Research from JSPS and MEXT, Health Sciences Research Grants from the Ministry of Health, Labour and Welfare of Japan, grants of the Virtual Research Institute of Aging of Nippon Boehringer Ingelheim, Mochida Medical and Pharmaceutical Research Foundation and a grant from Japan Orthopaedics and Traumatology Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.T. (taka.cell@tmd.ac.jp) or T.T. (tostakai@idac.tohoku.ac.jp).

Bacterial disease resistance in *Arabidopsis* through flagellin perception

Cyril Zipfel¹*, Silke Robatzek¹*, Lionel Navarro², Edward J. Oakeley¹, Jonathan D. G. Jones², Georg Felix¹* & Thomas Boller¹*

¹Friedrich Miescher Institute for Biomedical Research, PO Box 2543, CH-4002 Basel, Switzerland

²The Sainsbury Laboratory, John Innes Centre, Colney Lane, Norwich NR4 7UH, UK

* Present address: Botanical Institute, University of Basel, Hebelstrasse 1, CH-4056 Basel, Switzerland

Plants and animals recognize microbial invaders by detecting pathogen-associated molecular patterns (PAMPs)^{1–5} such as flagellin^{6–10}. However, the importance of flagellin perception for disease resistance has, until now, not been demonstrated^{7–11}. Here we show that treatment of plants with flg22, a peptide representing the elicitor-active epitope of flagellin⁶, induces the expression of numerous defence-related genes and triggers resistance to pathogenic bacteria in wild-type plants, but not in plants carrying mutations in the flagellin receptor gene *FLS2*. This induced resistance seems to be independent of salicylic acid, jasmonic acid and ethylene signalling. Wild-type and *fls2* mutants both display enhanced resistance when treated with crude bacterial extracts, even devoid of elicitor-active flagellin, indicating the existence of functional perception systems for PAMPs other than flagellin. Although *fls2* mutant plants are as susceptible as the wild type when bacteria are infiltrated into leaves, they are more susceptible to the pathogen *Pseudomonas syringae* pv. *tomato* DC3000 when it is sprayed on the leaf surface. Thus, flagellin perception restricts bacterial invasion, probably at an early step, and contributes to the plant's disease resistance.

In plants, disease resistance has been studied most thoroughly in cases that depend on the presence of specific resistance genes (*R* genes) conferring immunity to particular races of pathogens. The proteins encoded by *R* genes were shown to mediate specific recognition of factors specified by particular avirulence genes (*Avr* genes) in the pathogens^{12,13}. In addition to *R*-gene-related mechanisms, plants have broader, more basal perception systems for patterns characteristic for entire groups or classes of micro-organisms, so-called general elicitors¹⁴, which are conceptually equivalent to PAMPs^{1–5}. In contrast to *R*-gene-dependent defence, the responses to general elicitors do not always result in the cell death associated with the hypersensitive response, and the exact role of general elicitors in plant disease resistance is still unclear. PAMPs that act as general elicitors in plants include chitin¹⁵ and ergosterol¹⁶ from fungi, and flagellin⁶ and lipopolysaccharides¹⁷ from bacteria. Flagellin, the subunit building the filament of the bacterial flagellum, is also recognized as a PAMP in mammals, by way of the Toll-like receptor TLR5 (refs 9, 10). In *Arabidopsis*, perception of flagellin occurs by recognition of the most conserved domain in its amino terminus, represented by the peptide flg22 (ref. 6). Perception of this elicitor-active domain depends on the LRR-type receptor kinase FLS2 (flagellin sensing 2)⁸ and activates a downstream mitogen-activated protein kinase pathway, composed of *AtMEKK1*, *AtMKK4/AtMKK5* and *AtMPK3/AtMPK6* (ref. 7; the prefix *At* indicates *Arabidopsis thaliana*). flg22 induces numerous defence-related genes in *A. thaliana*, and the responses triggered by flg22 show great similarity to *R*-gene-mediated responses¹⁸.

Here we studied the role of flagellin perception in bacterial disease resistance. In a first step we extended a previous transcriptional analysis¹⁸ by comparing flg22-induced changes in intact wild-type and *fls2* mutant seedlings, using the full-genome Gene-Chip

ATH1 (Affymetrix) of *A. thaliana* (about 23,000 genes). After a 30-min treatment of wild-type seedlings with flg22, 966 genes were categorized as upregulated, 625 of them more than 2.5-fold; and 202 were categorized as downregulated, 35 of them more than 2.5-fold (Supplementary Fig. 1 and Supplementary Table 1). In seedlings of the flagellin-insensitive mutant *fls2-17*, carrying a point mutation in the kinase domain of FLS2 (G1064R)⁸, treatment with flg22 showed minor changes in six genes only (less than twofold changes). These genes do not belong to those regulated by flg22 in wild-type seedlings (Supplementary Table 1), indicating random fluctuations. This result demonstrates the validity of the criteria used to classify changes in the wild type as significant even below the 2.5-fold threshold, and it clearly shows that flagellin perception and signalling depend absolutely on the presence of a functional FLS2 receptor.

As well as a large group of genes with unknown functions (328 genes), a considerable number of the upregulated genes can be classified as being involved in signal perception (155 genes encoding receptor-like kinases (RLK) and *R* genes), signal transduction (145 genes), transcriptional regulation (87 genes), and potential antimicrobial action (29 genes) (Supplementary Table 2). Among the genes that are rapidly induced at the transcriptional level are the following genes (see MIPS database, <http://mips.gsf.de/prog/thal/db/index.html>): *FLS2* (At5g46330), *MEKK1* (At4g08500), *MKK4* (At1g51660), *MPK3* (At3g45640) and *WRKY22* (At4g01250). These genes encode elements that have previously been shown to be involved in the perception and transmission of the flg22 signal^{7,8}. A similar positive feedback regulation with transcriptional activation of the components involved in the perception and signalling has been reported for the innate immune response in *Drosophila*^{19,20}. Interestingly, our previous analysis of promoter sequences from flg22-induced genes revealed an over-representation of W-boxes¹⁸, that is, *cis*-elements, which confer the specific binding of WRKY

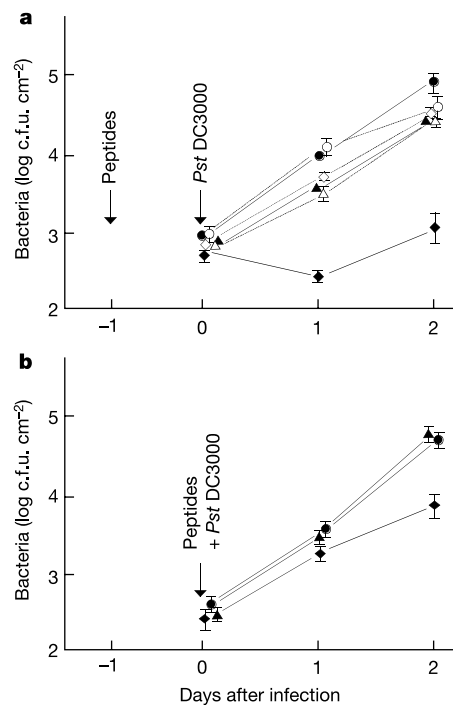


Figure 1 Treatment with flagellin limits *Pst* DC3000 growth. **a**, *Arabidopsis* wild-type Ler-0 (filled symbols) and *fls2-17* (open symbols) plants were pretreated for 24 h by leaf infiltration with 1 μ M flg22 (diamonds) or flg22 ^{Δ tum} (triangles). Subsequently, leaves were infected with 10⁵ c.f.u. ml⁻¹ *Pst* DC3000, and bacterial growth was assessed 1 and 2 days after infection. Control Ler-0 and *fls2-17* plants (circles) were not pretreated before bacterial infection. **b**, Ler-0 plants were infiltrated simultaneously with 1 μ M flg22 or flg22 ^{Δ tum} and 10⁵ c.f.u. ml⁻¹ *Pst* DC3000. Controls were treated with bacteria only. Results shown are means $\pm$ s.e.m. ($n = 8$).

transcription factors²¹. A similar over-representation of W-boxes was found when promoter sequences from all flg22-induced RLK genes were analysed (Supplementary Table 4). Because several WRKY factors are among the strongly induced genes after 30 min, it will be interesting to test whether a self-amplification system leads to even more pronounced induction of these genes after prolonged treatment with flg22. The induced RLKs and R genes constitute one-sixth of all upregulated genes (Supplementary Tables 2 and 3). With regard to the numbers of genes comprising these families in the *Arabidopsis* genome, this indicates an over-representation of 4.3-fold for RLKs and 3.8-fold for R genes, respectively. In summary, as well as the induction of numerous elements of the defence response, flagellin treatment seems to induce factors with an important function in the amplification of the signal and factors leading to an enhanced sensitivity of the plant to further stimuli sensing the presence of invading microorganisms. In particular, one could speculate that some of the induced RLKs and R genes might be involved in the recognition of other, as yet unidentified, PAMPs or Avr signals.

Transient overexpression of constitutively active MEKK1, MKK4, or wild-type WRKY29 resulted in reduced disease symptoms after treatment with *Pseudomonas syringae* pv. *maculicola* ES4326 or *Botrytis cinerea*⁷. To test whether direct induction of this signalling chain by flagellin leads to increased plant resistance, growth of pathogenic bacteria *in planta* was addressed after pretreatment of leaves with the flg22 elicitor. Wild-type and *fls2-17* mutant *Arabidopsis* plants were pretreated either with flg22 or the inactive analogue²² flg22^{A.tum} by leaf infiltration 1 day before challenge with pathogenic *Pseudomonas syringae* pv. *tomato* DC3000 (*Pst* DC3000) bacteria (Fig. 1a). In wild-type plants that received either no pretreatment (controls) or pretreatment with flg22^{A.tum}, bacteria multiplied at the same rate. However, bacterial growth was strongly decreased in plants pretreated with flg22 (about 100-fold difference at 2 days after infection). In *fls2-17* plants, pretreatment with flg22 did not lead to a decreased growth of bacteria (Fig. 1a). This result shows that the induction of resistance depends on a functional FLS2, and also that flg22 has no antimicrobial activity itself. Decreased bacterial growth was also observed, but less so, when flg22 peptide was applied concomitantly with the bacterial inoculum (Fig. 1b), indicating that exposure of elicitor-active flagellin present in the injected *Pst* DC3000 bacteria might be a limiting factor for efficient induction of the basal resistance, at least in the absence of R-gene-dependent detection of Avr factors.

Analysis of mutants and transgenic plants have revealed the importance of the salicylic acid, jasmonic acid and ethylene pathways in *Arabidopsis* resistance against pathogens²³. We tested the requirements of flg22-induced resistance for these previously identified defence signalling elements in plants mutated in *NPR1*,

EDS1, *SGT1*, *RAR1*, *ETR1*, *EIN2*, *JAR1*, *PAD2* or *PAD4*, or in plants overexpressing *NahG*. Whereas bacterial growth showed some accession-dependent and mutation-dependent variation, all mutants still exhibited a significant flg22-induced reduction in bacterial growth (Table 1). Thus, the genes tested are not required for flg22-induced resistance. It is surprising that *NPR1*, *EDS1* and *PAD4*, which are essential for salicylic-acid-mediated resistance²³ and are transcriptionally induced as rapidly as 30 min after treatment of seedlings with flg22 (Supplementary Table 2), are not involved in the observed flg22-induced resistance. Furthermore, *PR1*, whose induced expression through *NPR1* is a marker for salicylic-acid-mediated resistance²³, is activated 24 h after treatment with flg22 (ref. 24). However, our findings are consistent with the fact that flg22- and chitin-induced phosphorylation of the ankyrin-repeat protein *AtPhos43* was independent of *NPR1* and salicylic acid²⁵. flg22 activates the production of ethylene and triggers a rapid oxidative burst⁶. We therefore propose that flg22 induces the activation of the salicylic acid, jasmonic acid and ethylene pathways in parallel and that knocking out a single pathway alone does not abolish the induction of resistance. Alternatively, signalling elements that are as yet unknown, or the induction of antimicrobials, reactive oxygen species or cell wall reinforcement, might be responsible for inhibiting bacterial growth.

To test the involvement of PAMP perception systems other than flg22–FLS2 in resistance, we pretreated wild-type and mutant *fls2-17* plants with crude extracts obtained from *Pseudomonas syringae* pv. *syringae* (*Pss*), *Pst* DC3000 and *Agrobacterium tumefaciens*. All bacterial extracts—even that from *A. tumefaciens*, a species with an inactive flg22 peptide sequence²²—induced medium alkalization of *Arabidopsis* cell cultures (data not shown) and therefore showed clear elicitor activity. Pretreatment of plants with all the bacterial extracts resulted in a decreased growth of the pathogenic bacteria *Pst* DC3000 in comparison with that in control wild-type plants (Fig. 2). The same effect was also observed in *fls2-17* plants, but to a smaller extent in the case of a pretreatment with *Pst* DC3000 extracts. Thus, extracts from the tested bacteria contain at least one elicitor of resistance distinct from flagellin, and we propose that *Arabidopsis* has additional detection systems for these, as yet undefined, PAMPs.

Because flg22 perception induced disease resistance in plants, we tested whether plants lacking flagellin perception are more susceptible to pathogenic bacteria carrying elicitor-active flagellin. Inter-

Table 1 flg22-induced resistance in plants affected in salicylic acid, jasmonic acid and ethylene signalling

Line	Bacterial count (log c.f.u. cm ⁻²)	
	flg22 ^{A.tum}	flg22
Col-0	4.6 ± 0.2	3.4 ± 0.2
NahG	5.9 ± 0.01	4.7 ± 0.1
etr1-3	4.4 ± 0.3	2.7 ± 0.2
ein2-1	3.5 ± 0.25	2.7 ± 0.3
jar1-1	4.6 ± 0.15	3.5 ± 0.2
pad2-1	5.0 ± 0.1	3.9 ± 0.3
pad4-1	5.5 ± 0.2	3.6 ± 0.2
Ler-0	5.1 ± 0.1	3.5 ± 0.1
fls2-17	4.8 ± 0.25	4.9 ± 0.1
eds1-2	5.6 ± 0.05	3.7 ± 0.2
sgt1b-3	5.0 ± 0.2	3.6 ± 0.05
rar1-13	5.2 ± 0.2	3.5 ± 0.2
No-0	5.0 ± 0.2	2.6 ± 0.25
npr1-5	4.5 ± 0.15	2.2 ± 0.05

The following were pretreated for 24 h with 1 μM flg22 or flg22^{A.tum}: *Arabidopsis* transgenic *NahG* and mutants *etr1-3*, *ein2-1*, *jar1-1*, *pad2-1* and *pad4-1* in a Col-0 background; mutants *fls2-17*, *eds1-2*, *sgt1b-3* and *rar1-13* in a Ler-0 background; and mutant *npr1-5* in a No-0 background. Subsequent leaf infection with 10⁵ c.f.u. ml⁻¹ *Pst* DC3000 was performed, and bacteria were counted 2 days after infection as described in Methods.

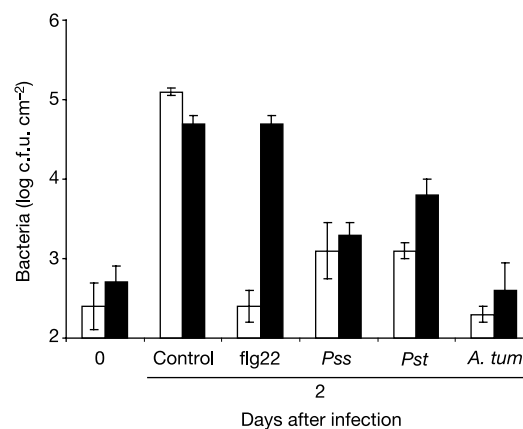


Figure 2 Treatment with different bacterial extracts limits subsequent growth of *Pst* DC3000 in Ler-0 and *fls2-17* plants. Ler-0 (open bars) and *fls2-17* (filled bars) plants were either left untreated, pretreated for 24 h with 1 μM flg22 or pretreated with one of the following bacterial extracts: *Pseudomonas syringae* pv. *syringae* (*Pss*) (3 mg ml⁻¹), *Pseudomonas syringae* pv. *tomato* DC3000 (*Pst* DC3000) (3 mg ml⁻¹) or *Agrobacterium tumefaciens* (*A. tum.*) (10 mg ml⁻¹). Subsequent leaf infection with 10⁵ c.f.u. ml⁻¹ *Pst* DC3000 was performed, and bacterial growth was assessed 2 days after infection. Results shown are means ± s.e.m. (*n* = 8).

estingly, *Pst* DC3000 bacteria infiltrated directly into the intercellular leaf space grew at the same rate in *fls2-17* as in wild-type plants (Fig. 1a) and caused the same visible disease symptoms within the first week after infection (data not shown). In addition, various non-pathogenic or avirulent strains of *Pseudomonas* and *Xanthomonas* grew at the same, restricted, rate in both the *fls2-17* and wild-type plants (data not shown). Under natural conditions, *Pst* DC3000 enters host plants, usually the leaves, through wounds or natural openings such as stomata, and then spreads and multiplies to high population densities in intercellular spaces²⁶. Thus, the infiltration of bacteria with a syringe might bypass the first steps of

the natural infection process, notably the steps of invasion and spreading that probably rely on flagella-based motility. Bacterial motility might not be important within the intercellular spaces, because non-motile mutants of *Pseudomonas syringae* pv. *phaseolicola* and *Xanthomonas campestris* pv. *malvacearum* were similar to the parental strains in their ability to grow after vacuum infiltration into bean and cotton leaves, respectively²⁶. We therefore infected *A. thaliana* plants by spraying *Pst* DC3000 bacteria onto leaf surfaces. Under these conditions, *fls2-17* plants showed a faster and more severe development of disease symptoms than wild-type plants (Fig. 3a). These stronger symptoms correlated with higher numbers of bacteria in *fls2-17* leaves (Fig. 3b), a difference that was particularly pronounced in younger leaves. Higher sensitivity of *fls2-17* mutants, compared to wild-type plants, was found in all of three independent experiments. Because *fls2-17* mutants originate from a population mutagenized with ethyl methane sulphonate, to exclude the possibility that the enhanced sensitivity of *fls2-17* was due to a genetic difference other than the one in the *FLS2* gene, we tested a second, independent, mutant affected in the *FLS2* gene in a Col-0 background. This mutant, carrying a T-DNA insertion in the promoter region abolishing expression of the *FLS2* gene (checked by reverse transcriptase polymerase chain reaction; data not shown), also showed enhanced sensitivity to *Pst* DC3000 in comparison with its wild-type Col-0 background (Supplementary Fig. 2). In addition, the ecotype Ws-0 presents a flagellin-insensitive phenotype^{22,24}, formerly attributed to a mutation in a hypothetical *FLS1* gene²⁴ but recently shown to be a natural *fls2* mutant carrying a point mutation that resulted in a stop codon in the kinase domain of *FLS2* (S.R., unpublished observations). In comparison with the accession Col-0, Ws-0 plants exhibited faster and more severe development of disease symptoms after being sprayed with *Pst* (data not shown). However, Ws-0 plants transformed with a functional *FLS2* gene, under the control of its native promoter sequence, acquired responsiveness to flg22 (Supplementary Fig. 3) and became less susceptible to *Pst* DC3000 (Fig. 3c), indicating that the natural deficiency in flagellin perception in the ecotype Ws-0 can be complemented with the wild-type *FLS2* gene.

Enhanced disease susceptibility to airborne infection with *Mycobacterium tuberculosis* or *M. avium* has been observed in knockout mice lacking a single Toll-like receptor (TLR2). However, a more drastic effect on susceptibility was observed in mice lacking MyD88, a signal adaptor protein thought to be required for transfer of signals coming from all TLRs²⁷. This indicates redundancy of the recognition process and the involvement of several TLRs in the innate immune system of animals. Interestingly, a common dominant TLR5 stop codon polymorphism abolishes flagellin signalling and is associated with susceptibility to Legionnaires' disease in humans²⁸. Similarly, the results presented above provide a first example that perception of a single general elicitor or PAMP makes a difference for plant defence. Although the sensing of flagellin by *FLS2* is an important initial checkpoint for controlling or restricting bacterial invasion in *Arabidopsis* leaves, it is not the only checkpoint; detection systems for additional bacterial PAMPs can be expected to have similar and complementary functions in controlling pathogen invasion at different steps of the infection process. The identification of these additional PAMP(s) and of the corresponding receptor(s) represents an exciting goal for the future. □

Methods

Plant material

All plants were grown at 20–21 °C with 65% humidity under light (about 100 $\mu\text{mol m}^{-2} \text{s}^{-1}$) in an 8 h light/16 h dark cycle in environment-controlled chambers. Plants aged 5–6 weeks were used for the infection experiments.

Flagellin treatment

Treatments with flg22 or flg22^{A.tum} were performed by pressure infiltration (needle-less syringes) of 1 μM peptide solution into the leaves. For each treatment, four to eight plant replicates were used, and each experiment was repeated at least twice.

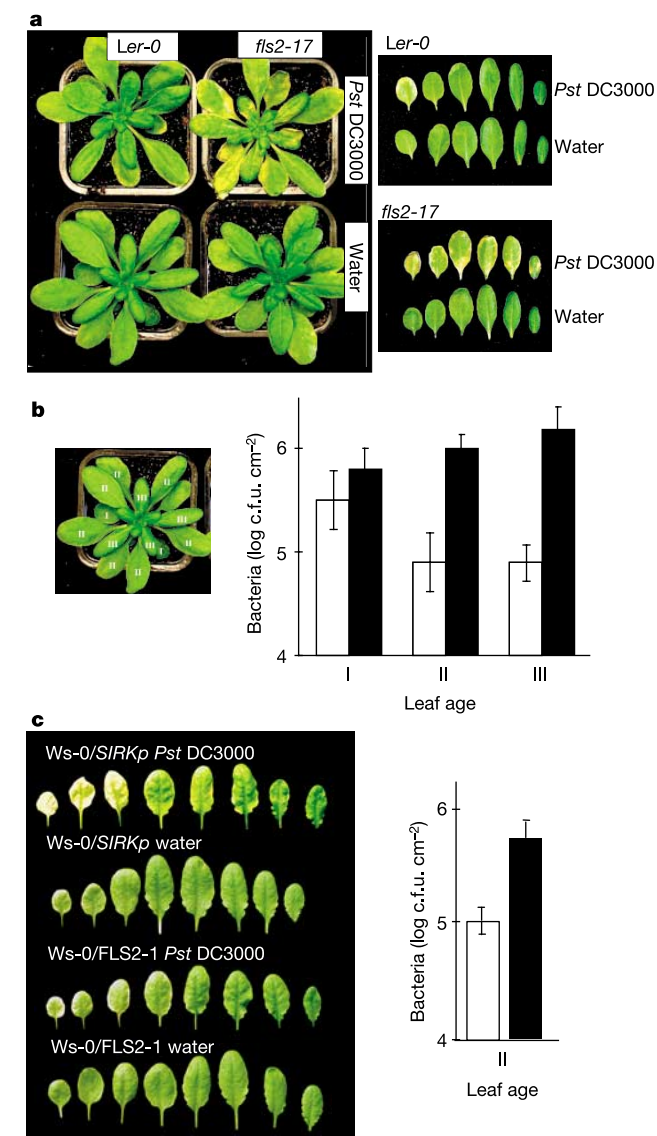


Figure 3 Bacterial disease resistance is determined by flagellin perception. **a**, A *FLS2* loss-of-function mutation, *fls2-17*, leads to enhanced disease susceptibility. Left: wild-type and *fls2-17* mutant plants were sprayed with *Pst* DC3000 bacteria or with water and photographed 4 days later. Right: symptoms after 4 days in a series of leaves of decreasing age. **b**, Number of *Pst* DC3000 bacteria extracted from wild-type (open bars) and *fls2-17* mutant plants (filled bars) 4 days after infection. Leaves were grouped by age as depicted on the left. **c**, A gain-of-function transgene of *FLS2* leads to decreased susceptibility in the accession Wassilewskaya (Ws-0), which lacks a functional *FLS2* gene. Ws-0 was stably transformed with *FLS2p::FLS2-3xmyc* (line *FLS2-1*; open bars), or *SIRKp::GUS* (line *SIRKp*; filled bars) as a control. Plants were sprayed with 5×10^8 c.f.u. ml^{-1} *Pst* DC3000, or water. Pictures were taken for a series of leaves of decreasing age (left), and bacteria were extracted and counted from leaves of age class II (right), 4 days after infection. Results are means $\pm$ s.e.m. ($n = 8$).

Bacterial growth assays

Bacterial strains used in this study were *Pseudomonas syringae* pv. *tomato* DC3000 (*Pst* DC3000), *Pst* DC3000 *AvrRpm1*, *Pst* DC3000 *AvrRps4*, *Pst* DC3000 *AvrRpt2*, *Pseudomonas syringae* pv. *tomato* DC3000 *HrpS*[−], *Pseudomonas syringae* pv. *syringae* (*Pss*), *Pseudomonas syringae* pv. *tabaci*, *Pseudomonas syringae* pv. *glyciniae*, *Pseudomonas syringae* pv. *phaseolicola*, *Pseudomonas fluorescens*, *Pseudomonas brassicacearum*, *Xanthomonas axonopodis* pv. *citri* and *Agrobacterium tumefaciens*. All strains were grown at 28 °C on King's B medium (40 g l^{−1} proteose, 20 g l^{−1} glycerol, 15 g l^{−1} agar) containing the appropriate antibiotics for selection. Syringe and spray inoculations, and bacterial growth in planta, were performed as described²⁹. In brief, for syringe inoculation, the bacteria were scraped off a fresh plate, resuspended in sterile water to 10⁵ colony-forming units (c.f.u.) ml^{−1}, and pressure-infiltrated into leaves with a needleless syringe. For spray inoculation, overnight *Pst* DC3000 cultures were collected, washed once and resuspended in sterile water. Plants were sprayed with a bacterial suspension containing 5 × 10⁸ c.f.u. ml^{−1} bacteria with 0.04% Silwet L-77 (Lehle Seeds). Leaves were harvested and surface sterilized (30 s in 70% ethanol, followed by 30 s in sterile distilled water) for the spray inoculation method. Leaf discs from two different leaves were ground in 10 mM MgCl₂ with a Microfuge tube glass pestle. After grinding of the tissue, the samples were thoroughly vortex-mixed and diluted 1:10 serially. Samples were finally plated on NYGA solid medium (5 g l^{−1} bactopectone, 3 g l^{−1} yeast extract, 20 ml l^{−1} glycerol, 15 g l^{−1} agar) supplemented with the appropriated antibiotic. Plates were placed at 28 °C for 2 days, after which the colony-forming units were counted.

Treatment with bacterial extracts

Extracts from *Pss* and *Pst* DC3000 were prepared as described⁶, freeze-dried and dissolved in water (3 mg ml^{−1}). *A. tumefaciens* bacteria were harvested by centrifugation, washed once with water and lysed by incubation in lysozyme solution (0.2 mg ml^{−1}) for 30 min at 37 °C and homogenization with a Polytron. The soluble supernatant was freeze-dried and redissolved in water (10 mg ml^{−1}).

Generation of transgenic plants

Arabidopsis thaliana Ws-0 plants were transformed with a *FLS2p::FLS2-3xmyc* construct. The *FLS2* promoter up to −988 base pairs was amplified by polymerase chain reaction (PCR) and introduced into the *EcoRI* and *HindIII* sites of pCambia 2300 (www.cambia.org.au), additionally adding *BamHI* and *KpnI* restriction sites upstream of the *HindIII* site. The *FLS2* gene triple Myc-tag fusion was amplified by PCR and cloned into the *BamHI* and *KpnI* sites of pCambia, and the construct was verified by sequencing. Stable transgenic lines were generated with the *A. tumefaciens*-mediated gene transfer procedure. Independent transformed plant pools were kept separate for the selection of independent transgenic lines based on their kanamycin resistance. Functional complementation of Ws-0 by *FLS2p::FLS2-3xmyc* was assayed with standard procedures²⁴. Expression of *FLS2-3xmyc* protein was confirmed by western blot analysis with anti-Myc antibodies (Supplementary Fig. 3). For control transformation, a *SIRKp::GUS* construct³⁰ was used. Plants of the T₂ generation were chosen for the bacterial spraying experiments.

Received 24 December 2003; accepted 11 March 2004; doi:10.1038/nature02485.

- Medzhitov, R. & Janeway, C. A. Jr Decoding the patterns of self and nonself by the innate immune system. *Science* **296**, 298–300 (2002).
- Takeda, K., Kaisho, T. & Akira, S. Toll-like receptors. *Annu. Rev. Immunol.* **21**, 335–376 (2003).
- Tzou, P., De Gregorio, E. & Lemaitre, B. How *Drosophila* combats microbial infection: a model to study innate immunity and host–pathogen interactions. *Curr. Opin. Microbiol.* **5**, 102–110 (2002).
- Kurz, C. L. & Ewbank, J. J. *Caenorhabditis elegans*: an emerging genetic model for the study of innate immunity. *Nature Rev. Genet.* **4**, 380–390 (2003).
- Aballay, A., Drenkard, E., Hilbun, L. R. & Ausubel, F. M. *Caenorhabditis elegans* innate immune response triggered by *Salmonella enterica* requires intact LPS and is mediated by a MAPK signaling pathway. *Curr. Biol.* **13**, 47–52 (2003).
- Felix, G., Duran, J. D., Volk, S. & Boller, T. Plants have a sensitive perception system for the most conserved domain of bacterial flagellin. *Plant J.* **18**, 265–276 (1999).
- Asai, T. *et al.* MAP kinase signalling cascade in *Arabidopsis* innate immunity. *Nature* **415**, 977–983 (2002).
- Gómez-Gómez, L. & Boller, T. *FLS2*: an LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in *Arabidopsis*. *Mol. Cell* **5**, 1003–1011 (2000).
- Smith, K. D. *et al.* Toll-like receptor 5 recognizes a conserved site on flagellin required for protofilament formation and bacterial motility. *Nature Immunol.* **4**, 1247–1253 (2003).
- Hayashi, F. *et al.* The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* **410**, 1099–1103 (2001).
- Nürnberger, T. & Brunner, F. Innate immunity in plants and animals: emerging parallels between the recognition of general elicitors and pathogen-associated molecular patterns. *Curr. Opin. Plant Biol.* **5**, 318–324 (2002).
- Dangl, J. L. & Jones, J. D. Plant pathogens and integrated defence responses to infection. *Nature* **411**, 826–833 (2001).
- Staskawicz, B. J., Mudgett, M. B., Dangl, J. L. & Galan, J. E. Common and contrasting themes of plant and animal diseases. *Science* **292**, 2285–2289 (2001).
- Boller, T. Chemoperception of microbial signals in plant cells. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **46**, 189–214 (1995).
- Felix, G., Regenass, M. & Boller, T. Specific perception of subnanomolar concentrations of chitin fragments by tomato cells. Induction of extracellular alkalization, changes in protein phosphorylation, and establishment of a refractory state. *Plant J.* **4**, 307–316 (1993).
- Granado, J., Felix, G. & Boller, T. Perception of fungal sterols in plants: subnanomolar concentrations of ergosterol elicit extracellular alkalization in tomato cells. *Plant Physiol.* **107**, 485–490 (1995).
- Dow, M., Newman, M. A. & von Roepenack, E. The induction and modulation of plant defense responses by bacterial lipopolysaccharides. *Annu. Rev. Phytopathol.* **38**, 241–261 (2000).
- Navarro, L. *et al.* The transcriptional innate immune response to flg22: interplay and overlap with Avr

gene-dependent defence responses and bacterial pathogenesis. *Plant Physiol.* (in the press).

- De Gregorio, E., Spellman, P. T., Rubin, G. M. & Lemaitre, B. Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays. *Proc. Natl Acad. Sci. USA* **98**, 12590–12595 (2001).
- Irving, P. *et al.* A genome-wide analysis of immune responses in *Drosophila*. *Proc. Natl Acad. Sci. USA* **98**, 15119–15124 (2001).
- Eulgem, T., Rushton, P. J., Robatzek, S. & Somssich, I. E. The WRKY superfamily of plant transcription factors. *Trends Plant Sci.* **5**, 199–206 (2000).
- Bauer, Z., Gómez-Gómez, L., Boller, T. & Felix, G. Sensitivity of different ecotypes and mutants of *Arabidopsis thaliana* toward the bacterial elicitor flagellin correlates with the presence of receptor-binding sites. *J. Biol. Chem.* **276**, 45669–45676 (2001).
- Hammond-Kosack, K. E. & Parker, J. E. Deciphering plant–pathogen communication: fresh perspectives for molecular resistance breeding. *Curr. Opin. Biotechnol.* **14**, 177–193 (2003).
- Gómez-Gómez, L., Felix, G. & Boller, T. A single locus determines sensitivity to bacterial flagellin in *Arabidopsis thaliana*. *Plant J.* **18**, 277–284 (1999).
- Peck, S. C. *et al.* Directed proteomics identifies a plant-specific protein rapidly phosphorylated in response to bacterial and fungal elicitors. *Plant Cell* **13**, 1467–1475 (2001).
- Beattie, G. A. & Lindow, S. E. The secret life of foliar bacterial pathogens on leaves. *Annu. Rev. Phytopathol.* **33**, 145–172 (1995).
- Feng, C. G. *et al.* Mice lacking myeloid differentiation factor 88 display profound defects in host resistance and immune responses to *Mycobacterium avium* infection not exhibited by Toll-like receptor 2 (TLR2)- and TLR4-deficient animals. *J. Immunol.* **171**, 4758–4764 (2003).
- Hawn, T. R. *et al.* A common dominant TLR5 stop codon polymorphism abolishes flagellin signaling and is associated with susceptibility to Legionnaires' disease. *J. Exp. Med.* **198**, 1563–1572 (2003).
- Katagiri, F., Thilmony, R. & He, S. Y. in *The Arabidopsis Book* (eds Somerville, C. R. & Meyerowitz, E. M.) doi:10.1199/tab.0039, <http://www.aspb.org/publications/arabidopsis> (American Society of Plant Biologists, Rockville, Maryland, 30 September 2002).
- Robatzek, S. & Somssich, I. E. Targets of AtWRKY6 regulation during plant senescence and pathogen defense. *Genes Dev.* **16**, 1139–1149 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank H. Angliker for help with the microarray procedure; R. Ulm and D. Chinchilla for critically reading the manuscript; C. Molteni and B. Thürig for technical help; the Nottingham *Arabidopsis* Stock Centre for *etr1-3*, *ein2-1*, *jar1-1*, *pad2-1* and *pad4-1* seeds; the Torrey Mesa Research Institute for the SAIL_691-C4 line; J. Parker for *eds1-2*, *sgt1b-3* and *rar1-13* seeds; W. Achouak for the *Pseudomonas brassicacearum* strain; S. Y. He for *Pseudomonas syringae* pv. *tomato* DC3000 *HrpS*[−] mutant strain; and N. Kraus and H. Sierotzky for the plant growing facilities. This work was supported by the Novartis Research Foundation, by the Gatsby Charitable Foundation and by a grant from the Swiss National Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.B. (thomas.boller@unibas.ch).

Cdc42 and mDia3 regulate microtubule attachment to kinetochores

Shingo Yasuda^{1,2}, Fabian Oceguera-Yanez¹, Takayuki Kato^{1*}, Muneo Okamoto¹, Shigenobu Yonemura³, Yasuhiko Terada⁴, Toshimasa Ishizaki¹ & Shuh Narumiya¹

¹Department of Pharmacology and ²Horizontal Medical Research Organization, Kyoto University Faculty of Medicine, Kyoto 606-8501, Japan

³Laboratory for Cellular Morphogenesis, RIKEN Center for Developmental Biology, Kobe 650-0047, Japan

⁴Department of Genetics, Cell Biology and Development, University of Minnesota, Minnesota 55455, USA

* Present address: Department of Physiology, Osaka City University Medical School, Osaka 545-8585, Japan

During mitosis, the mitotic spindle, a bipolar structure composed of microtubules (MTs) and associated motor proteins^{1,2}, segregates sister chromatids to daughter cells. Initially some MTs emanating from one centrosome attach to the kinetochore at the centromere of one of the duplicated chromosomes. This attachment allows rapid poleward movement of the bound chromosome. Subsequent attachment of the sister kinetochore to MTs

growing from the other centrosome results in the bi-orientation of the chromosome, in which interactions between kinetochores and the plus ends of MTs are formed and stabilized². These processes ensure alignment of chromosomes during metaphase and their correct segregation during anaphase. Although many proteins constituting the kinetochore have been identified and extensively studied, the signalling responsible for MT capture and stabilization is unclear^{1,2}. Small GTPases of the Rho family regulate cell morphogenesis by organizing the actin cytoskeleton and regulating MT alignment and stabilization³. We now show that one member of this family, Cdc42, and its effector, mDia3, regulate MT attachment to kinetochores.

Although Rho participates in cytokinesis^{3,4}, whether Rho GTPases contribute to nuclear division remains unclear. To examine the possibility that Rho GTPases do take part, we synchronized HeLa cells in prometaphase of mitosis using nocodazole, and incubated the nocodazole-arrested cells with *Clostridium difficile* toxin B, which inactivates all Rho GTPases including Rho, Rac and Cdc42⁵. Cells were able to progress through mitosis when nocodazole was removed. Staining for DNA, MTs and a kinetochore protein, CENP-A², revealed that toxin B treatment markedly prevented chromosome alignment at metaphase (Fig. 1a). Even 90 min after nocodazole removal, when control cells had progressed to anaphase and telophase, the chromosomes of many toxin-treated

cells were widely scattered, with several chromosomes clustering near the poles despite most cells retaining a bipolar spindle. Scattered chromosomes were often mono-oriented (Fig. 1a, insets 1 and 2) or orientated with an apparent lack of MT attachment (Fig. 1a, inset 3). Staining for Mad-2, a mitotic spindle-checkpoint protein², verified the loss of MT attachment to these chromosomes (Fig. 1b), which was further confirmed by electron microscopy (Supplementary Information). The cells with misaligned chromosomes decreased gradually from 90% of the population at 60 min to 50% at 180 min (Fig. 1c). Video microscopy (Supplementary Movies 1 and 2) revealed that the chromosomes of toxin-B-treated cells moved without congression during the entire observation period of 250 min, and that some exited to interphase without a distinct anaphase to produce aberrant tetraploid nuclei. This progression to interphase was accompanied by cyclin B degradation (Supplementary Information), suggesting that the toxin B treatment affected the spindle-checkpoint mechanism.

The above findings implicated a Rho family GTPase in MT

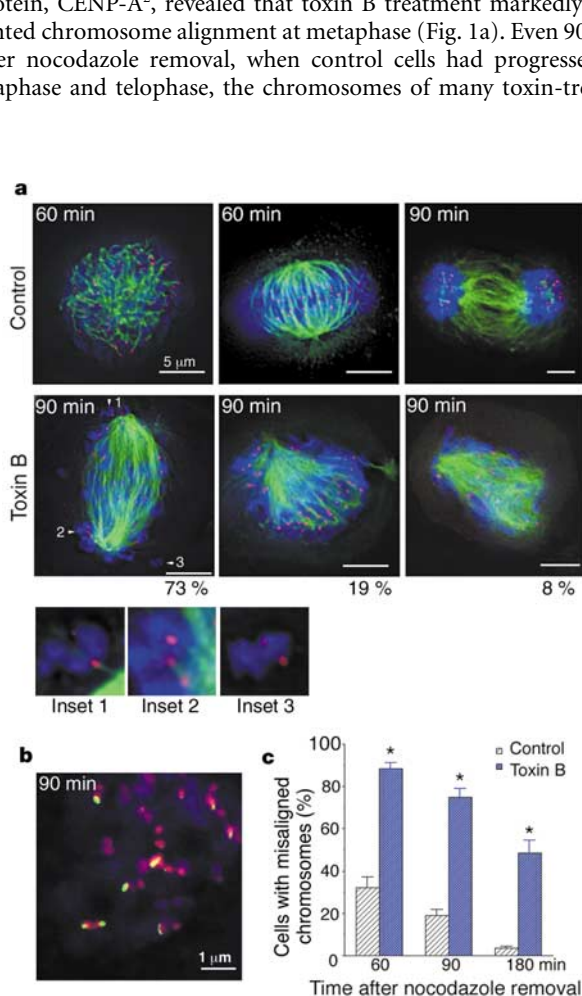


Figure 1 Effects of toxin B on mitosis. **a**, Chromosome misalignment in toxin-B-treated cells. Control or toxin-B-treated HeLa cells were fixed at the times indicated after nocodazole removal and stained. β -tubulin, green; CENP-A, red; DNA, blue. The numbers below the lower panels (**a**) indicate the percentage of cells exhibiting (from left to right) the bipolar, monopolar and multipolar spindle, respectively ($n = 300$). The insets show the parts of the image indicated by the numbered arrows in the above figure. **b**, MAD2 staining of toxin-B-treated cells 90 min after removal of nocodazole. CENP-A, red; MAD2, green; DNA, blue. **c**, Percentage of cells with misaligned chromosomes. Data are means $\pm$ s.e.m. from 100 cells at each time point ($n = 5$). * $P < 0.05$ versus control experiment.

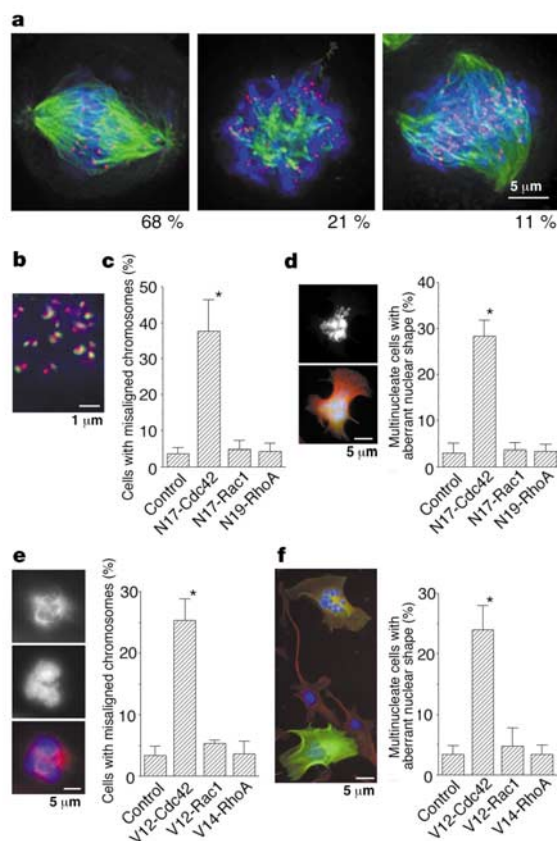


Figure 2 Effects of Rho GTPase mutants on mitosis. **a, b**, Chromosome misalignment (**a**) and MAD2 staining (**b**) of N17-Cdc42 expressing cells. Staining is the same as in Fig. 1. The numbers shown in **a** indicate the percentage of cells with abnormal mitosis exhibiting (from left to right) the bipolar, monopolar and multipolar spindle, respectively ($n = 98$). **c, d**, Percentages of cells with chromosome misalignment (**c**) or aberrant multiple nuclei (**d**) among cells expressing green fluorescent protein (GFP) alone (control) or the GFP fusion of N17-Cdc42, N17-Rac1 or N19-RhoA in mitosis or interphase. The left panel in **d** shows an N17-Cdc42 expressing multinucleate cell (β -tubulin, red; DNA, blue). **e, f**, Chromosome misalignment (**e**) and production of multinucleate cells (**f**) by expression of V12-Cdc42 in NIH 3T3 cells. The percentages of cells with misaligned chromosomes or multiple nuclei among cells expressing GFP alone (control) or GFP fusion of V12-Cdc42, V12-Rac1 or V14-RhoA were analysed at mitosis (**e**) or interphase (**f**). Left panels show images of cells expressing V12-Cdc42 and stained for β -tubulin and DNA as in **d**. Data in **c-f** are means $\pm$ s.e.m. of 100 cells for each condition ($n \geq 3$, each). * $P < 0.05$ versus control cells.

attachment and chromosome alignment during mitosis, therefore we examined the effects of dominant-negative mutants of Rho GTPases on mitosis. HeLa cells were synchronized at the beginning of S phase using a double-thymidine block and transfected with complementary DNAs for the mutant proteins during the middle of the second block. Most cells entered mitosis 12 h after thymidine removal. Cells expressing the dominant-negative N17-Cdc42 mutant exhibited chromosome misalignment similar to that in toxin-B-treated cells (Fig. 2a). Again, Mad2 localized to kinetochores of the misaligned chromosomes, indicating that MTs failed to attach to the kinetochores (Fig. 2b). These abnormalities were detected in >35% of cells expressing N17-Cdc42 (Fig. 2c). By contrast, the proportion of cells with these phenotypes was less than 5% in control cells transfected with the empty vector and those expressing the dominant-negative N19-RhoA or N17-Rac1 mutants. Sixteen hours after thymidine removal, multinucleate cells with macro- and micronuclei and nuclear bridges were frequently observed among interphase cells expressing N17-Cdc42 (Fig. 2d), indicating that the chromosome misalignment in these cells led to missegregation of chromosomes. By contrast, and consistent with previous reports^{4,6}, expression of N19-RhoA caused cytokinesis failure, producing cells with two nuclei of equal size (data not shown). No phenotype was apparent in cells expressing N17-Rac1 (data not shown). Chromosome misalignment and generation of multinucleate cells were also induced by expression of a dominant-active mutant of Cdc42 (V12-Cdc42) (Fig. 2e, f), but not of Rac1 or RhoA. These results suggest that Cdc42 is essential for

MT-mediated chromosome alignment in mammalian cells, and that either inhibition or enhancement of its activity causes chromosome misalignment, prometaphase delay and chromosome missegregation.

We next attempted to identify the downstream target of Cdc42 in this action. A candidate was the Rho effector mDia, a mammalian homologue of *Drosophila* diaphanous and a member of the formin-homology (FH) family of proteins, which are conserved from yeast to mammals⁷. This family of proteins induces actin polymerization and promotes the alignment and stabilization of MTs^{8–10}. There are three mDia isoforms (mDia1⁸, mDia2¹¹ and mDia3/hDia2¹²), all of which contain a Rho-binding domain (RBD) and the highly conserved FH1, FH2 and FH3 regions (Fig. 3a).

Using a yeast two-hybrid assay, we found that the FH2 region of mDia isoforms binds to CENP-A^{2,15} and a different region of mDia binds to another centromere-kinetochore protein, heterochromatin protein (HP) 1¹⁶ (Fig. 3a and Supplementary Information). We also found that among the three isoforms only mDia3 can bind to Cdc42, as well as to RhoA and Rac1, in a GTP-dependent manner (Fig. 3b and Supplementary Information). These results suggest that mDia3 is a target of Cdc42 at the kinetochore. Consistently, we found mDia3 in the prekinetochore complex precipitated with an anti-CENP-A antibody¹⁷ from nuclear digests but not in the control pellets (Fig. 3c).

An immunofluorescence study using a specific anti-mDia3 antibody (Supplementary Information) detected mDia3 signals as punctae in the nucleus of interphase cells that colocalized with

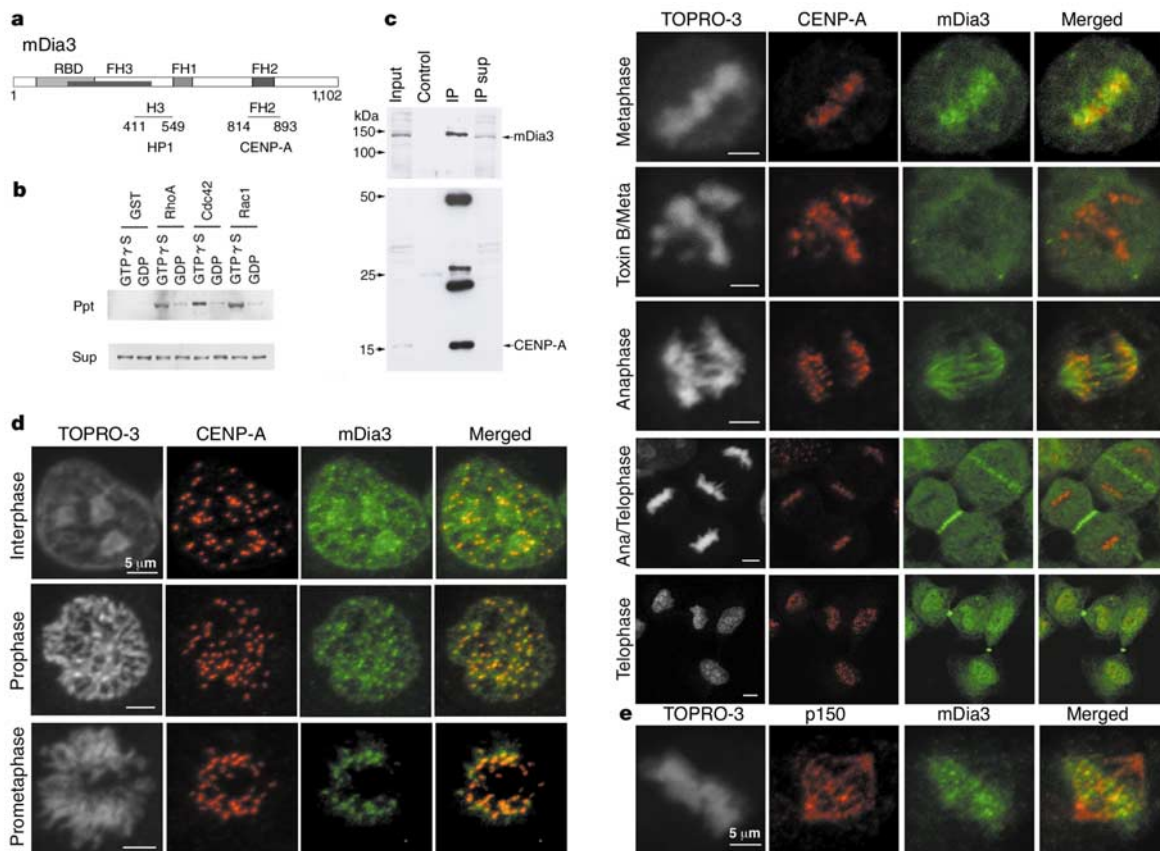


Figure 3 Interaction of mDia3 and CENP-A. **a**, Domain structure of mDia3. The regions interacting with HP-1 and CENP-A are indicated. **b**, GTP-dependent binding of mDia3 to Rho GTPases. An *in vitro* precipitation assay between the RBD of mDia3 and Rho GTPases bound to GDP or GTP-γS is shown. **c**, Co-precipitation of mDia3 and CENP-A in the prekinetochore complex. Nuclear digests were precipitated by control beads or anti-

CENP-A antibody (IP) and probed with antibodies to mDia3 (above) or CENP-A (below). **d**, Localization of mDia3 in mitosis. HeLa cells were fixed at the indicated phases and stained for mDia3 (green), CENP-A (red) and DNA (TOPRO-3). Toxin B/meta, a toxin B-treated mitotic cell. **e**, Co-staining for mDia3 (green) and p150^{Glued} (red) in mitotic cells.

HP-1 signals (Supplementary Information) and partially overlapped with those of CENP-A (Fig. 3d). More of the mDia3 signals overlapped with CENP-A during prophase nuclei, and they were present together at kinetochores of condensed chromosomes in prometaphase cells (Fig. 3d). During metaphase, mDia3 signals were still associated with CENP-A (Fig. 3d) and were juxtaposed with those of p150^{Glued}, a component of the dynein-dynein complex that anchors MTs to the kinetochore² (Fig. 3e). Toxin B treatment abolished mDia3 signals at the kinetochore without affecting CENP-A localization (Fig. 3d), indicating that endogenous Cdc42 localizes mDia3 at the kinetochore in metaphase chromosomes. By contrast, mDia3 localization was not affected by active Cdc42 expression (data not shown), suggesting that dominant-active Cdc42 causes chromosome misalignment not by displacement of mDia3 from the kinetochore but instead by activating mDia3 homogenously in the cell, thereby decreasing the significance of endogenous Cdc42. During anaphase, some mDia3 signals moved poleward with CENP-A, but some migrated to both polar and midzone spindle MTs. The signals in the midzone became condensed during late anaphase and accumulated in the midbody in telophase. Thus, a subpopulation of mDia3 behaved like a chromosomal passenger protein¹⁸. By contrast, mDia1 was localized predominantly in the cytoplasm of interphase cells, and was enriched in the polar spindle MTs and the cell cortex of mitotic cells, as previously described^{8,14} (Supplementary Information).

We next investigated the role of mDia3 during mitosis. We transfected HeLa cells with a small interfering RNA (siRNA) duplex targeted to a region of mDia3 messenger RNA. As controls, we used siRNA with a scrambled sequence or targeted to mDia1 mRNA¹⁹. Treatment with siRNA for mDia1 and mDia3 resulted in a marked decrease in the respective proteins as detected by immunoblotting (Fig. 4a) and confirmed by immunofluorescence (Supplementary Information). Examination 36 h after transfection revealed that the mitotic index of the mDia3 RNAi cells increased significantly, to 25%, whereas control cells had a mitotic index of 8 to 9% (Fig. 4b). In this mitotic population of mDia3 RNAi cells, a marked increase

in the percentage of cells in prometaphase and a concomitant decrease in those in metaphase, anaphase or telophase was found (Fig. 4c).

Video microscopy of these prometaphase-arrested cells found that most of them had established bipolar spindles, but many chromosomes remained scattered at the poles (Fig. 4d; Supplementary Movies 3–5). This is the same phenotype found in cells treated with toxin B or expressing N17-Cdc42. To identify the mechanism underlying this abnormality, we stained for EB1, a plus-end MT-binding protein that is important for spindle MT stabilization². Consistent with previous reports^{20,21}, some EB1 signals in control cells colocalized with those of CENP-A on metaphase chromosomes. However, they barely overlapped with CENP-A signals in cells treated with either siRNA for mDia3 or toxin B (Fig. 4e). In addition to these findings, we also noted that expression of a dominant-active form of mDia3 in mitotic NIH 3T3 cells induced chromosomal misalignment and resulted in production of multinucleate cells (Supplementary Information).

We have shown involvement of Cdc42 and its effector mDia3 in MT-mediated chromosome alignment. This role for Cdc42, although unexpected, is consistent with several observations. For example, although not explicitly stated, expression of dominant-negative forms of ECT2 or MgcRacGAP, a putative activator and downregulator, respectively, of Rho GTPases in mitosis, induced the formation of tetraploid nuclei with abnormal shape^{22,23}. Homozygous disruption of the mouse *MgcRacGAP* gene resulted in death of day 3.5 to 4 embryos owing to the failure of chromosomal segregation²⁴. Using a pull-down assay, we found that Cdc42 was activated at metaphase, and this activation was regulated by Ect2 and MgcRacGAP (unpublished observation). Roles for Cdc42, mDia and homologues, often in conjunction with EB1 and its binding partner, adenomatous polyposis coli protein (APC), in the regulation of MT alignment have been reported previously, in yeast²⁵ and mammalian cells^{9,10,26}. Notably, mDia1 induces MT alignment and stabilization at the cell cortex through EB1 in mammalian cells^{9,10}. Therefore, it appears that similar mecha-

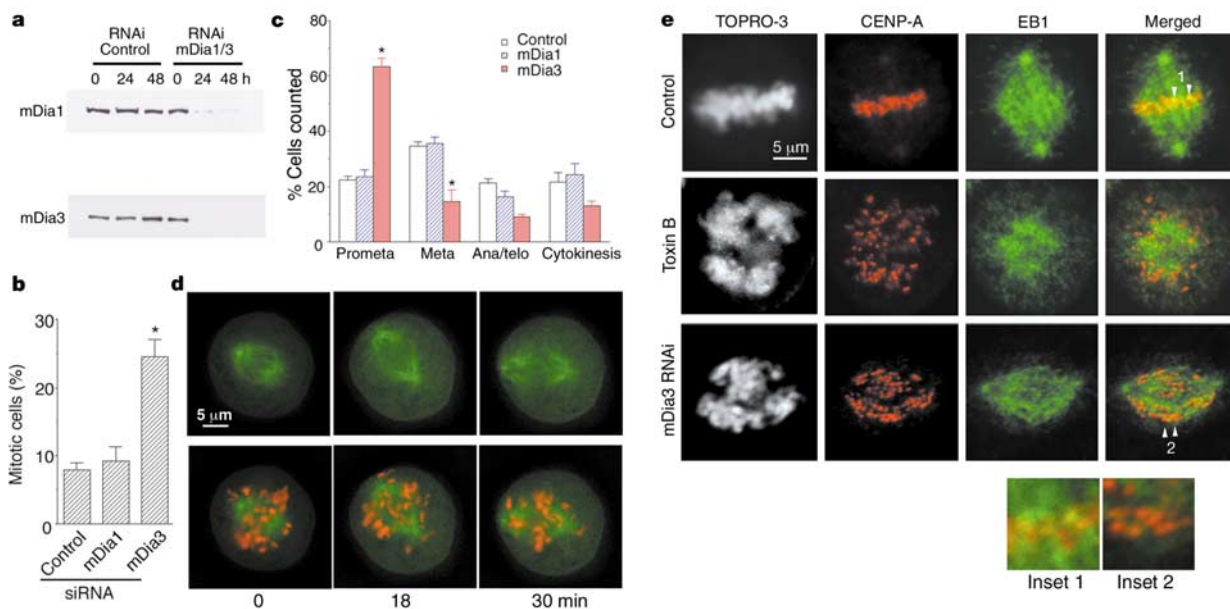


Figure 4 Effects of depletion of mDia3 on mitosis. **a**, Depletion of mDia1 and mDia3 by RNA interference. Immunoblots of cells subjected to RNAi with antibodies to mDia1 and mDia3 are shown. **b**, **c**, Increase of the mitotic index and prometaphase arrest in mDia3 RNAi cells. The mitotic index of 300 cells subjected to indicated RNAi in three independent experiments (**b**), and the proportion of the cells in each phase of mitosis (**c**) are shown as means $\pm$ s.e.m. * $P < 0.05$ versus the control cells. **d**, Video microscopy. Cells with

GFP-EB1-labelled MTs (green) and with dsRed2-histone H2B-k-labelled chromosomes (red) were subjected to RNAi for mDia3, and progression through mitosis was monitored by video-microscopy. Images for GFP-EB1 alone are shown above. (Movies available online as Supplementary Movies 3–5.) **e**, Loss of EB1 signals at kinetochores in cells treated with toxin B or subjected to RNAi for mDia3. The insets show parts of the images indicated by the numbered arrows in the figures.

nisms operate during MT attachment to the cell cortex and kinetochores. APC and EB1 are also important for MT attachment to kinetochores^{2,27}. MTs attach first to one of the kinetochore pair of the chromosomes, causing translocation of chromosomes to the corresponding pole. MT attachment then occurs in a bi-oriented manner and is stabilized^{2,28}. In cells with impaired Cdc42–mDia3 signalling, many chromosomes accumulate near each pole and are either mono-oriented or without MT attachment. These findings indicate that the Cdc42–mDia3 signalling is not involved in the initial attachment of MTs to kinetochores but plays an important role in subsequent, stable bi-orientated MT attachment for proper chromosome alignment and segregation. Whether such a role for the Cdc42–mDia3 signalling is limited to mammalian cells remains to be seen. □

Methods

Materials

Sources and construction of plasmid DNAs are described in Supplementary Information. The mDia1 siRNA was described previously¹⁹, and the mDia3 siRNA corresponded to nucleotides 172 to 190 of the mRNA relative to the translation initiation codon.

Cell culture and transfection

HeLa cells and NIH 3T3 cells were maintained at 37 °C in Dulbecco's modified Eagle's medium (GIBCO-BRL) supplemented with 10% fetal bovine serum and penicillin–streptomycin. Cells were synchronized in S phase by the double-thymidine block protocol (two cycles of 15 h culture with thymidine and 9 h culture free of thymidine); enrichment of mitotic cells was achieved by nocodazole treatment after release from the thymidine block²⁹. When transfection was performed during the second thymidine block, HeLa cells were washed 9 h after the initiation of the second thymidine block, transfected in Opti-MEM with the use of Lipofectamine Plus (GIBCO-BRL) for 2 h, and cultured again in the original culture medium for 4 h in the continued presence of thymidine before thymidine removal. Expression in NIH3T3 cells was carried out by microinjection of plasmid DNA immediately after thymidine removal.

Microscopy

For immunofluorescence, cells were fixed in methanol at –20 °C. For kinetochore staining, cells were permeabilized for 5 min with 0.005% digitonin in a solution containing 110 mM potassium acetate, 20 mM HEPES (pH 7.3), 2 mM magnesium acetate, 0.5 mM EGTA and 2 mM dithiothreitol before fixation. Immunostaining was performed as described previously⁸. Primary antibodies included those against mDia3 (goat, 1:50 dilution; Santa Cruz Biotechnology), mDia1¹³ (rabbit; 1:200 dilution), α -tubulin (DM1A, mouse, 1:100 dilution; Sigma), β -tubulin (TUB2.1, mouse, 1:100 dilution; Sigma), MAD2 (rabbit, 1:500 dilution; COVANCE), CENP-A (human CREST serum, 1:1,000 dilution), p150^{Glued} (mouse, 1:200 dilution; Transduction Laboratories) and EB1 (mouse, 1:200 dilution; Transduction Laboratories). Immune complexes were detected with appropriate secondary antibodies labelled with either Alexa488, Alexa594 or Alexa633 (Molecular Probes). DNA was stained with Hoechst 33342 (1 μ g ml^{–1}), DAPI or TOPRO-3 (Molecular Probes). Cells were examined with a Zeiss LSM510 confocal imaging system or a DeltaVision microscope system. Electron microscopy was carried out as described previously³⁰.

Assays for protein interaction and staining

Two-hybrid assays were performed as described previously¹³ with the plasmids described above. The L40 yeast strain was cotransfected with pBTM and either pVP16 or pACT2 vectors, and protein interaction was detected by staining with the β -galactosidase substrate X-gal. Coprecipitation ('pull-down') assays were performed using recombinant glutathione-S-transferase–Rho GTPases and thioredoxin-tagged RBDs as described previously¹³. For chromatin immunoprecipitation, 5×10^7 HeLa cells were harvested and homogenized in 10 mM Tris.HCl containing 10 mM NaCl, 3 mM MgCl₂ and 0.5% NP-40. Isolated nuclei were suspended in nuclei-washing buffer and were digested for 1 h at 37 °C with 80 U micrococcal nuclease¹⁷. The digests were centrifuged at 10,000g for 10 min, and the supernatant was subjected to precipitation with mouse monoclonal anti-CENP-A antibody¹⁷. The immunoprecipitates were recovered and probed with antibodies to CENP-A and mDia3. Western blot analysis was performed using an anti-cyclin B1 antibody (clone CB169, Upstate biotechnology) or antibodies to mDia1 and mDia3 as described previously⁸.

Received 11 December 2003; accepted 1 March 2004; doi:10.1038/nature02452.

- Scholey, J. M., Brust-Masher, I. & Mogilner, A. Cell division. *Nature* **422**, 746–752 (2003).
- Cleveland, D. W., Mao, Y. & Sullivan, K. F. Centromeres and kinetochores: from epigenetics to mitotic

checkpoint signaling. *Cell* **112**, 407–421 (2003).

- Etienne-Manneville, S. & Hall, A. Rho GTPases in cell biology. *Nature* **420**, 629–635 (2002).
- Mabuchi, I. *et al.* A rho-like protein is involved in the organisation of the contractile ring in dividing sand dollar eggs. *Zygotes* **1**, 325–331 (1993).
- Just, I. *et al.* Glucosylation of Rho proteins by *Clostridium difficile* toxin B. *Nature* **375**, 500–503 (1995).
- Eda, M. *et al.* Rho-dependent transfer of Citron-kinase to the cleavage furrow of dividing cells. *J. Cell Sci.* **114**, 3273–3284 (2001).
- Wasserman, S. FH proteins as cytoskeletal organizers. *Trends Cell Biol.* **8**, 111–115 (1998).
- Watanabe, N. *et al.* p140mDia, a mammalian homolog of *Drosophila* diaphanous, is a target protein for Rho small GTPase and is a ligand for profilin. *EMBO J.* **16**, 3044–3056 (1997).
- Ishizaki, T., Morishima, Y., Furuyashiki, T., Kato, T. & Narumiya, S. Coordination of microtubules and actin cytoskeleton by a Rho effector, mDia1. *Nature Cell Biol.* **3**, 8–14 (2001).
- Gundersen, G. G. Evolutionary conservation of microtubule-capture mechanisms. *Nature Rev. Mol. Cell Biol.* **3**, 296–304 (2002).
- Alberts, A. S., Bouquin, N., Johnston, L. H. & Treisman, R. Analysis of RhoA-binding proteins reveals an interaction domain conserved in heterotrimeric G protein β subunits and the yeast response regulator protein Skn7. *J. Biol. Chem.* **273**, 8616–8622 (1998).
- Bione, S. *et al.* A human homologue of the *Drosophila melanogaster* diaphanous gene is disrupted in a patient with premature ovarian failure: evidence for conserved function in oogenesis and implications for human sterility. *Am. J. Hum. Genet.* **62**, 533–541 (1998).
- Watanabe, N. *et al.* Cooperation between mDia1 and ROCK in Rho-induced actin reorganization. *Nature Cell Biol.* **1**, 136–143 (1999).
- Kato, T. *et al.* Localization of a mammalian homolog of Diaphanous, mDia1, to the mitotic spindle in HeLa cells. *J. Cell Sci.* **114**, 775–784 (2001).
- Goshima, G., Kiyomitsu, T., Yoda, K. & Yanagida, M. Human centromere chromatin protein hMis12, essential for equal segregation, is independent of CENP-A loading pathway. *J. Cell Biol.* **160**, 25–39 (2003).
- Kellum, R. HP1 complexes and heterochromatin assembly. *Curr. Top. Microbiol. Immunol.* **274**, 53–77 (2003).
- Ando, S. *et al.* CENP-A, -B, and -C chromatin complex that contains the I-type α -satellite array constitutes the prekinetochore in HeLa cells. *Mol. Cell Biol.* **22**, 2229–2241 (2002).
- Adams, R. R., Carmena, M. & Earnshaw, W. C. Chromosomal passengers and the (aurora) ABCs of mitosis. *Trends Cell Biol.* **11**, 49–54 (2001).
- Arakawa, Y. *et al.* Control of axon elongation via an SDF-1 α /Rho/mDia pathway in cultured cerebellar granule neurons. *J. Cell Biol.* **161**, 381–391 (2003).
- Tirnauer, J. S., Canman, J. C., Salmon, E. D. & Mitchison, T. J. EB1 targets to kinetochores with attached, polymerizing microtubules. *Mol. Biol. Cell* **13**, 4308–4316 (2002).
- Anand, S., Penrhyn-Lowe, S. & Venkataraman, A. R. AURORA-A amplification overrides the mitotic spindle assembly checkpoint, inducing resistance to Taxol. *Cancer Cell* **3**, 51–62 (2003).
- Tatsumoto, T., Xie, X., Blumenthal, R., Okamoto, I. & Miki, T. Human ECT2 is an exchange factor for Rho GTPases, phosphorylated in G2/M phases, and involved in cytokinesis. *J. Cell Biol.* **147**, 921–928 (1999).
- Hirose, K. *et al.* MgcRacGAP is involved in cytokinesis through associating with mitotic spindle and midbody. *J. Biol. Chem.* **276**, 5821–5828 (2001).
- Van de Putte, T. *et al.* Mice with a homozygous gene trap vector insertion in *mgcRac GAP* die during pre-implantation development. *Mech. Dev.* **102**, 33–44 (2001).
- Rodriguez, O. C. *et al.* Conserved microtubule-actin interactions in cell movement and morphogenesis. *Nature Cell Biol.* **5**, 599–609 (2003).
- Etienne-Manneville, S. & Hall, A. Cdc42 regulates GSK-3 β and adenomatous polyposis coli to control cell polarity. *Nature* **421**, 753–756 (2003).
- Pellman, D. Cancer. A CINTillating new job for the APC tumor suppressor. *Science* **291**, 2555–2556 (2001).
- Yao, X. *et al.* CENP-E forms a link between attachment of spindle microtubules to kinetochores and the mitotic checkpoint. *Nature Cell Biol.* **2**, 484–491 (2000).
- Andreassen, P. R., Palmer, D. K., Wener, M. H. & Marholis, R. L. Telophase disc: a new mammalian mitotic organelle that bisects telophase cells with a possible function in cytokinesis. *J. Cell Sci.* **99**, 523–534 (1991).
- Maiato, H. *et al.* Human CLASP1 is an outer kinetochore component that regulates spindle microtubule dynamics. *Cell* **113**, 891–904 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank K. Aktories for *Clostridium difficile* toxin B, Y. Kiyosue for pQB125-X β -tubulin, N. Mimori for CREST serum, S. Tsukita for the use of Delta-Vision system, T. Kiyomitsu, T. Tsuji, Y. Arakawa, J. Monypenny and N. Watanabe for advice, and M. Yanagida for discussion. This work was supported in part by Grants-in-aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan and from the Ministry of Health, Labour and Welfare of Japan.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.N. (snaru@four.med.kyoto-u.ac.jp).

Data's future shock

Databases are having to move with the times as people expect more from them than simple data storage and retrieval. Steve Buckingham investigates.

Today's databases are faced with a moving target. There is so much more information available, and the type of data being collated is changing. Decisions have to be made: how should a database be restructured to accommodate the new information? What do we do with the old data? Are they compatible with the new results, or must they be hived off and archived in some way?

Take protein structure databases. Not long ago it was enough to submit a set of atomic coordinates that described a protein's structure. Now, such databases are expected to store 'meta-data' as well — how the protein was produced and purified, and how its structure was solved. And the rise in high-throughput projects will make yet greater demands.

But administrators of public databases are keeping pace with the changes. When the Protein Data Bank (PDB) was set up at Brookhaven National Laboratory in 1971 it held seven structures. Today it has more than 22,000 X-ray and NMR protein and peptide structures. And whereas it was once the exclusive haunt of crystallographers, the PDB is



PDB

Adding value: the PDB team now curates more than 22,000 protein structures.

now regularly used by biologists of all kinds.

The success of databases such as the PDB in keeping pace with these changes is due largely to careful planning. "We are preparing for three challenges for the future," says Helen Berman, the PDB's director: "The effects of structural genomics, the need for better

storage of macromolecular complex data, and internationalization. There's a lot of change in the pipeline." Berman expects structural genomics to double the number of data items attached to each protein submitted to the PDB. The database is also ready for the increasing interest in macromolecular

EXPLORING THE PUBLIC DOMAIN

A vast body of annotated and linked data is available in the public databases. But how do you find the database that best fits your needs? One place to start is the supplement produced as the first issue of each year by the journal *Nucleic Acids Research*, which is free online.

Alternatively, you could plunge straight into one of the large, general-purpose, bioinformatics databases such as the Ensembl Genome Browser (EGB) or the National Center for Biotechnology Information's Entrez portal. Most are now so closely integrated with more specialized databases that

navigation through a collection of databases is all but seamless. Careful design makes them powerful tools even in the hands of a novice, yet it is easy to progress to more sophisticated use. Simple search-text boxes take a term through a selection of databases, and various options control the amount and type of data returned. Help buttons, along with links that provide simple explanations of each term, are never far away.

For genomics, the EGB is a popular port of entry. This collaborative effort between the Sanger Institute, the European Bioinformatics Institute (EBI) and the European Molecular Biology Laboratory provides automated annotation of human, mouse, rat, fugu, zebrafish, mosquito, fruitfly and nematode genomes. The homepage contains a link to an 'Ensembl tour' and 'worked examples'. Users can search for a term across species for protein, disease or mRNA, or can follow links to a page dedicated to each species.

PEP, a database of Predictions for Entire Proteomes, is the result of a sophisticated analysis of proteomes from over 60 species. The work of Burkhard Rost's bioinformatics group at Columbia University, New York, PEP primarily contains open reading frames (ORFs) along with predicted structural domains detected within the ORFs. PEP can be searched online either directly or through the EBI. Alternatively, the entire PEP database can be downloaded — if you have space.

Some commercial companies offer free online access to their own high-quality databases to academic researchers, such as the MendelBase database of structural and functional protein information from Array Genetics of Newtown, Connecticut.

S.B.

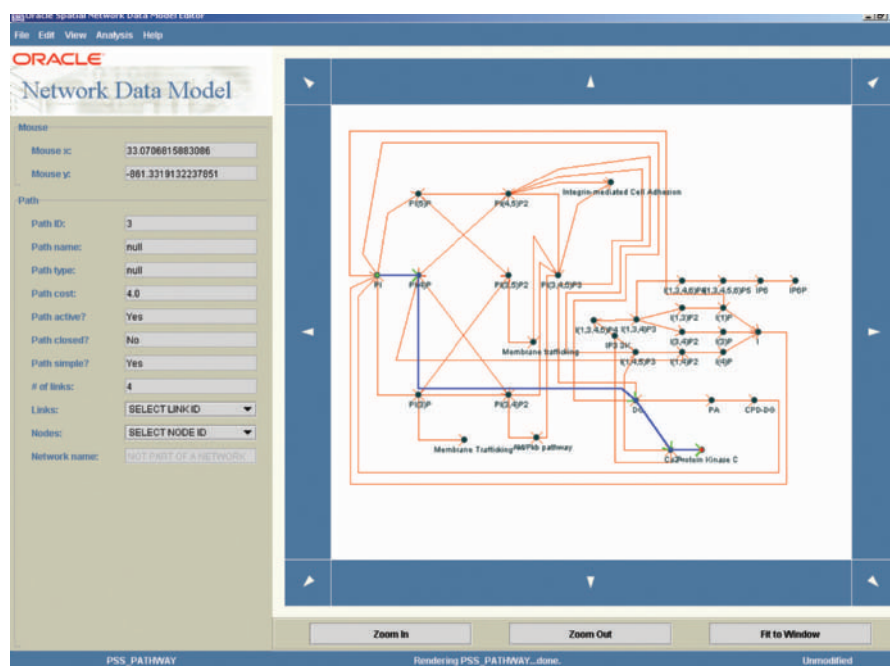


Free for all: from genomes to proteomes.

complexes, such as viral assemblies and ribosomes. Integrating these new data with the existing store is not easy but, as Berman says, "As science moves, the database must move with it." Beth Smith, director of solutions development at IBM in Somers, New York, agrees. "Annotation is going to lead to a huge increase in volume data," she says. "As medicine moves towards targeted treatment as a result of genomic approaches, we will see the rising need for high-performance computers and storage hardware. We aim to stay ahead of that capacity."

Planners have had to develop strict but extensible standards. In the case of the PDB, these took the form of the 'macromolecular dictionary' format. This has some 1,700 terms that not only define a protein's structure but also how that structure was solved. It encapsulates details of data types used in crystallographic descriptions, as well as the relationships between those data. And it is expandable — new entries are made according to strict procedures, so that new data types will always be fully integrated with older data.

As databases have changed, so has the software underpinning them. Database software company Oracle, of Redwood Shores, California, already has some 75–80% share of the general database market worldwide, and two years ago it turned its attention to the lucrative life-sciences market. "We are an opportunistic organization," says Susie Stephens, a senior life-sciences product manager at Oracle. "We see that the life-science database area is a substantial and sustainable business."



Pathway to knowledge: Oracle's Spatial Network Data Editor.

Database software is striving to meet the demands of this market. Users want access to distributed data with full integration of different data types. Technologies embedded in Oracle's database software, for example, allow a query to be run across distributed databases of different types, including non-Oracle and flat-file databases. Users also want to manage large quantities of data, and to be able to adjust the capacity of their hardware

to the size of their database and the demands placed upon it. Oracle's answer is Real Application Cluster (RAC) technology, which makes it easy to add new servers, or nodes, to an existing set of servers on the fly, in response to demand, and without having to reconfigure the whole database.

Oracle's new database release, 10g, is its first to incorporate features specifically geared to the life sciences, such as pattern-

BUYING INTO THE KNOWLEDGE GAME

Despite the impressive public databases, commercial ones can sometimes offer added value and convenience. They typically incorporate at least some information that is not available in the public domain, and have also done much of the hard work of annotating sequences and collating genomic and proteomic information.

Iconix Pharmaceuticals of Mountain View, California, for example, offers the DrugMatrix chemogenomics database and informatics system, which integrates public-domain chemical data with thousands of results from its experiments on the effects of known drugs and related compounds on gene expression and cell biology. DrugMatrix can help predict the effects of a test compound on gene expression and identify compounds that have similar effects to those in the database.

In its Discovery Knowledge database suite, MDL in San Leandro, California, offers two chemical databases, CrossFire Beilstein and CrossFire Gmelin, covering organic and inorganic chemistry, respectively. These databases are installed on a local server for access through proprietary browser software. MDL also offers Biopendium from Inpharmatica in London, which enables researchers to identify known drug targets and select related proteins in a range of experimental model systems. It uses comparisons of sequence, structure and ligand interactions, presented via a interactive alignment editor, ligand-interaction viewer and three-dimensional structure viewer. MDL's Discovery Gate structure-searchable literature information resource, combining 17 chemistry-related databases, is now also available on an academic licence.

Bringing a variety of information together in one convenient package is the selling point for commercial databases. For smaller research departments, data purchasing can fill big gaps in research capability. Buying databases can, for example, effectively bring high-throughput approaches within their reach. BioMax Informatics of Martinsried, Germany, for instance, offers reasonably priced subscription access to an annotated human genome database. The most recent release also includes the mouse genome and is integrated with the ProChart protein-interaction database from peptide-synthesis company AxCell, in Newtown, Pennsylvania.

Available online through an academic or commercial licence, the LifeSeq Foundation database from Incyte in Palo Alto, California, provides manually annotated and highly collated data on the sequence, expression and function of some 18,000 complete human genes and many more expressed sequence tags, including proprietary data not available in public databases. Each gene or gene fragment in LifeSeq Foundation is annotated with comprehensive functional information, including its relevance to disease. The database also contains information on the tissues in which a gene is expressed, related genes in the human genome, counterparts in model organisms, and known mutations. Incyte's ZooQuest database extends LifeSeq Foundation to cover mouse, rat, monkey and dog, and its Proteome Bioknowledge Library complements these databases with manually curated information gleaned from the literature on protein function and interaction for humans and selected model organisms.

S.B.

recognition functions, built-in BLAST search, and embedded machine-learning algorithms such as support vector machines for the analysis of microarray gene-expression data, for example. There are also built-in routines that allow searches using 'regular expressions' — complex word-pattern matching — that complement the powers of the favourite bioinformatics programming language Perl.

But the real power of databases is the ability to unearth patterns hidden across different types of data. For this, a database must be able to query widely different types of information in a common format. "Databases are becoming more capable of doing analysis through different data types and allowing integration of different types of data," says Jacek Myczkowski, Oracle's vice-president for life sciences and data-mining technologies. For example, patterns of gene expression from patients with different forms of a disorder can be stored in a relational database table, along with written clinical notes. Algorithms such as Oracle's support vector machines can then be used to build models using these two data types to identify the gene-expression patterns that are the most reliable markers of each disease profile.

Even data mining of unstructured text has seen some astonishing advances. Oracle

Text will read a document and provide an intelligent summary. "A document identified as being about cars, for example, can mention Audi and BMW and not even mention the word 'car'," says Stephens. "Oracle Text routines can extract the theme of a document like this, and can identify its subject matter".

Working together

The integration of databases is a priority if the full potential of the genomics revolution is to

be realized. "There is a clear trend today to get all these databases working together," says Joe Donahue, US president of LION Bioscience in Cambridge, Massachusetts. "Databases have always had cross-references to each other, but now we can search across them all at once."

To do this, each database needs to know something of the hidden workings of the others, such as the names of its database fields and what sort of data those fields contain. These were once closely guarded secrets, but things are changing. "The attitude only a few years ago was, 'my database is better than yours,'" says Berman. "But now everyone realizes that there is far too much work to do.

We have to marshal our resources."

This openness is good news, but will databases ever merge seamlessly? Myczkowski is pessimistic: "There can be no permanent standards because of the pace of change in the data." Steve Gardner at text-database company BioWisdom of Cambridge, UK, agrees. "You will never get people to adhere to standards enough to semantically integrate databases," he says. "There have been strides made in the technology to map data structures together using rule-based or ad hoc strategies, but all these systems fall down because they need rules that link fields from one database to another." But it is not all gloom. Run a query against your favourite protein at the European Bioinformatics Institute (EBI) website, and you'll see it run seamlessly against a host of diverse databases housed at separate institutions and developed by different authors with different uses in view.

Database maintenance

For most research groups, however, setting up their own database of any significant size or complexity is not easy. Even when finished, a database needs to be updated regularly, the new data have to be parsed, indexed and stored, and special software often has to be developed. So, despite the desirability of an in-house, home-made database, the cost of maintaining it can be prohibitive for a small research group.

Paris-based Gene-IT aims to fill this gap in the market. Later this year the firm will launch its GenomeCast automatic



Joe Donahue: different databases must work together.

GETTING THE MEANING

Although a relative newcomer to bioinformatics, ontologies have already attracted commercial interest. BioWisdom of Cambridge, UK, supplies ontologies in various fields. "Life science R&D poses a multidimensional problem," says Steve Gardner, BioWisdom's chief technical officer. "The problem is being able to communicate the information to a user interested not just in a molecule, but also in the context surrounding that molecule." BioWisdom currently offers more than 10 million distinct concepts linked by over 100 million relationships.

BioWisdom can also assist researchers to develop their own ontology. The first task is to build a database framework to encapsulate it. An additional framework embeds methods to normalize the incoming data, so that an entity is recognized despite having different names in different data sources. This is not easy: the sedative diazepam, for example, has some 197 synonyms.

Good ontology software can even help the researcher develop new hypotheses. "We have inferencing programs that draw together different concepts," says Gardner. "If one ontology says that COX2 is expressed in synovocytes, and another says that synovocytes are implicated in rheumatoid arthritis, the inferencing program would suggest that COX2 may be implicated in rheumatoid arthritis."

The output of an ontology is a graph: a representation of the relationships between concepts. Once a graph has been

generated, users can then bring their experience to bear. For example, they can exclude types of information on the strength of the evidence. "We call this a semantic lens," says Gardner. "You pass this lens over the data and it filters them out like a polarizing filter. This makes a new graph that lets you highlight the interactions that are interesting to you." BioWisdom's system has a hierarchical family of relationships: the protein-to-protein class, for example, has 400 potential relationships (such as 'interacts with',

'upregulates' and 'activates'). Thus, ontologies allow the user to search using one key term by resolving the meaning of that term, and then searching against it.

A taste of how ontologies work is provided by the public-domain Genome Ontology (GO) Browser, which gives free access to the genome ontologies developed by the GO Consortium. Three ontologies have been developed: molecular function, biological process and a cellular component. Using the Ensembl GO browser, the user can find the Ensembl genes that have been mapped to these ontologies. The search term is presented at the centre of a 'mind map'. Clicking on a 'child' or 'parent' term will produce a new Ensembl GO report centred on that term. The genes found are listed, along with links to different types of views of each gene and its chromosomal location. The ontologies can be also searched directly, with the results showing the connections between the terms.



Steve Gardner: linking concepts.

database-update system, complementing its GenomeQuest sequence-search system, which has recently been adopted by the European Patent Office. GenomeCast is aimed at both small labs and large drug firms. GenomeCast will automatically perform regular database upkeep without human intervention. As new data are posted on public databases, the program will aggregate them online, annotate them and combine them into a common format native to the GenomeQuest search engine. This eliminates the need for continual monitoring by a database administrator, and is rather like having someone doing your database administration for you remotely. Ron Ranauro, general manager of Gene-IT, believes that GenomeCast is following a trend in the database field. "The game has shifted away from providing curated scientific content towards delivering increasing amounts of data in real time along with the best tools at a reasonable cost and within a reasonable IT framework," he says.

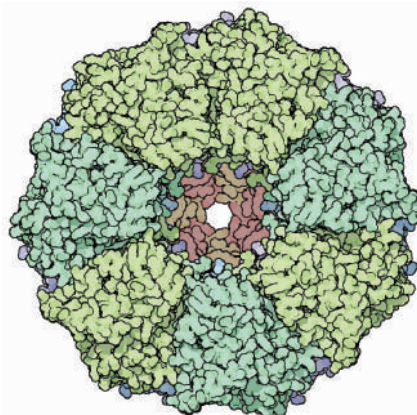
Let's talk

When it comes to databases talking to each other, there are five broad approaches to the problem of relating data entries from different databases: rules-based approaches, data warehousing, search optimizers, federation and ontologies.

Rules-based systems operate by specifying explicitly the relationship between different fields in different databases. This approach relies on records of the same object in different databases sharing some identifier, or cross-reference. With genes, for example, this might be the GenBank accession code. LION's SRS technology is the rules-based system that underlies the interoperability of the European Molecular Biology Laboratory, Wellcome and Sanger Institute databases, along with nearly 20,000 other commercial and academic databases. Users of SRS have made the details of their database structure, along with the parsers of their data, available to the SRS system. So, as more institutions use SRS, it can integrate more data, creating an ever-widening circle of interoperability.

SRS is bundled into a coordinated package, SRS Evolution, along with SRS Relational (for accessing relational structures), SRS 3D (for integrating protein structures), and other modules that assist data downloading and expression analysis. SRS technology also underlies LION's DiscoveryCenter software, aimed principally at the drug-discovery market, a package that allows a single point of access to a number of databases with integrated analysis applications.

LION's collaborations with the pharmaceutical industry are resulting in new software solutions. "We are expecting a big surge in the field of pharmacogenomics," says Simon



What's in a name? This chaperonin complex can be tracked through the databases by its PDB ID 1aon.

Baulah of LION. "Our customers are starting to use SRS to integrate patient data and gene-expression data in improving personalized medicine." A collaboration between LION and the Cambridge-based UK Human Genome Mapping Project has resulted in integration into SRS of the recently developed EMBOS query suite, a free set of bioinformatics applications that rivals the performance of the commercial Wisconsin package from Accelrys of San Diego, California.

An alternative approach to database unification is warehousing — making a local copy of data drawn from diverse sources and then forging them into a common format in a unified, specialized database. This approach can be expensive in time and money, but commercial warehousing solutions from companies such as Iconix Pharmaceuticals in Mountain View, California, Incyte in Palo Alto, California, and Gene-IT could be one answer. Alternatively, tools such as DS SeqStore from Accelrys make in-house data warehousing easier. Using a client/server architecture built on Oracle, this program helps to set up a secure database complete with analysis tools and a complete GenBank, GenPept, SWISS-PROT with SP-TrEMBL distribution. Its open architecture makes it comparatively easy to adapt the design to the user's corporate or lab needs.

Another way to search across different databases is to use query-optimizing systems. These use a battery of strategies to recast the query until the best results are returned from the databases. This is the approach behind the Discoverylink system from IBM, which uses a set of 'wrappers' to adapt a query to the databases questioned. "Discoverylink allows optimization of searches across a number of different databases with diverse formats, but the user is presented with only one interface," says Smith.

In the federation approach, member databases agree to represent data in a certain way, so that no adjustment has to be made to har-

monize them. An example is FEDORA (Federation of Research Assets), a federation of six special HTTP servers, released by Metaphorics of La Jolla, California. Data are federated into a knowledge base comprising a set of hyperlinks between synonyms and near synonyms, which permits sophisticated data mining. The current FEDORA cluster includes Empath, a server for metabolic pathway information, Planet, a server for protein-ligand information and WDI, a server for the World Drug Index.

An understanding approach

Human languages are rich in synonyms and subtleties of vocabulary. But this very richness makes cross-database searching a hit-and-miss affair. This looks set to change with the introduction of new techniques, such as ontologies, that are beginning to grapple directly with semantic complexity. Ontologies are networks of objects, their properties, and their relations to one another. They try to tackle the meanings of words, rather than just treating them as strings. This means more than just reducing linguistic complexity: ontologies actively exploit it.

Ontologies allow a concept in one resource (or database) to be mapped onto a concept from another. For example, an ontology might contain a representation of the fact that muscarinic acetylcholine receptors are G-protein-coupled receptors. This would allow a search for G-protein-coupled receptors across different databases, even though the search term 'G-protein-coupled receptor' might not occur in all of them.

In an ontology, the core concept (a gene for example) at the centre is connected to related concepts (such as coexpressed genes, proteins, diseases, tissues or compounds) which in turn are connected to yet more concepts (see 'Getting the meaning', opposite). But the links in ontologies are much more versatile than just a simple line, they can express relationships such as 'BINDS-TO' or 'IS-EXPRESSED-IN'. Ontologies are dynamic maps of information space and, like sourdough, once you have got it started, you can go on adding to it.

Ontologies are still in their infancy. But if they deliver what they promise, their contribution to making new insights could be enormous.

Steve Buckingham is a neuroscientist at the Medical Research Council's Functional Genomics Unit at Oxford University, UK.

PDB

◆ www.rcsb.org/pdb

Ensembl GO browser

◆ www.ensembl.org/Homo_sapiens/goview

Nucleic Acids Research database issue 2004

◆ nar.oupjournals.org/content/vol32/suppl_1

Predictions for Entire Proteomes

◆ cubic.bioc.columbia.edu/pep

D. GOODSELL/SCRIPPS RES. INST.

Company	Products/activity	Location	URL
Databases			
Ardais	Biomaterials and Information for Genomic Research (BIGR) library for genomics-based biomedical research	Lexington, Massachusetts	www.ardais.com
BIOBASE	TRANSFAC family of databases; analysis tools for gene expression, promoters and signalling pathways; contract bioinformatics	Wolfenbüttel, Germany	www.biobase.de
Biomax Informatics	Annotated human-genome database; customized data management	Martinsried, Germany	www.biomax.de
BioWisdom	Text search, and pharmacology and oncology information databases	Cambridge, UK	www.biowisdom.com
Celera Genomics	Celera Discovery System for accessing the Celera annotated genomes databases; bioinformatics services	Rockville, Maryland	www.celera.com
Cengent	Protein and protein-structure databases; computational proteomics	San Diego, California	www.strubix.com
Compugen	GenCarta annotated human genome, transcriptome and proteome database	Tel-Aviv, Israel	www.cgen.com
GeneLogic	Gene-expression databases and software for drug discovery	Gaithersburg, Maryland	www.GeneLogic.com
Iconix	DrugMatrix databases and software platform for chemogenomics research	Mountain View, California	www.iconixpharm.com
Incyte Genomics	LifeSeq and ZooSeq ESTdatabases; BioKnowledge Library protein-information databases; bioinformatics software	Palo Alto, California	www.incyte.com
Lexicon Genetics	Gene-knockout and gene-function databases and bioinformatics for drug discovery	The Woodlands, Texas	www.lexgen.com
LifeSpan BioSciences	Gene-expression and protein-localization databases and data-mining tools	Seattle, Washington	www.lsbio.com
MDL	CrossFile Beilstein, CrossFire Gmelin and other chemistry-related databases	San Leandro, California	www.mdl.com
PubGene	PubGene public-access and commercial gene databases, and analysis software	Oslo, Norway	www.pubgene.com
Database systems, data-management and data-mining software			
Accelrys	GCG Wisconsin Package; MacVector for sequence analysis; Discovery Studio suite for Windows for database mining, genomics and proteomics	San Diego, California	www.accelrys.com
Affymetrix	GeneChip microarray systems and gene-expression analysis tools	Santa Clara, California	www.affymetrix.com
ApoCom Genomics	Desktop bioinformatics tools for gene prediction (GrailEXP, GeneHound) and gene-expression analysis (Excavator)	Knoxville, Tennessee	www.apocom.com
Array Genetics	Online proteomics information database; bioinformatics tools for genomics, proteomics and microarray analysis	Newtown, Connecticut	www.arraygenetics.com
Biocomputing Platforms	Data-management systems for genotype and phenotype data	Espoo, Finland	www.biocomputing.fi
Bioinformatics Solutions	Desktop bioinformatics software: PatternHunter for DNA/protein analysis and PROSPECT for protein-structure prediction	Waterloo, Canada	www.bioinformaticsolutions.com
BioTools	Analysis software for gene and protein sequences, and chromatograms	Edmonton, Canada	www.biotools.com
Cognia	Bioinformatics software, including BIOBASE software and databases	New York, New York	www.cognia.com
Curagen	GeneScape portal for genome-analysis tools	Branford, Connecticut	www.curagen.com
DNASTAR	Lasergene sequence analysis software for Macintosh and Windows; GenVision genomic data visualization tool	Madison, Wisconsin	www.dnastar.com
DECODON	Software for 2D-gel analysis and information storage	Greifswald, Germany	www.decodon.de
Entigen	BioNavigator platform for sequence and genome analysis	Sunnyvale, California	www.entigen.com
GATC Biotech	Accelrys, DNASTAR and other bioinformatics software; DNA sequencing	Constance, Germany	www.gatc-biotech.com
GeneBio	Commercial access to SWISS-PROT databases; bioinformatics software	Geneva, Switzerland	www.genebio.com
Gene Codes	Sequencher sequence assembly and analysis software	Ann Arbor, Michigan	www.genecodes.com
Gene-IT	Universal software for sequence database management, comparison, clustering and results analysis	Le Chesnay, France	www.gene-it.com
Genomatix Software	Bibliosphere for literature searches and data mining; software for genome annotation, analysis and retrieval of promoters, and genome analysis	Munich, Germany	www.genomatix.de
Genomic Solutions	Genomics and proteomics bioinformatics tools	Ann Arbor, Michigan	www.genomicsolutions.com
Geospiza	Finch servers and tools for sequence assembly, analysis and data management	Seattle, Washington	www.geospiza.com
Hitachi Software Engineering	DNASIS desktop bioinformatics software for DNA sequence assembly and analysis, and analysis of microarray data	Yokohama, Japan	www.hitachi-sk.co.jp/english/index.html
IBM	DiscoveryLink platform for database integration; data-management systems	Somers, New York	www.ibm.com/solutions/lifesciences
Inpharmatica	Biopendium and CeleraEdition Biopendium proteome annotation resources; PharmaCarta chemogenomics informatics platform	London, UK	www.inpharmatica.com
InforMax	Vector bioinformatics software for sequence, genome and microarray data; Vector NTI for Macintosh; LabShare for data storage and management	Bethesda, Maryland	www.informaxinc.com
Iobion Informatics	GeneTraffic microarray data-management and analysis software	La Jolla, California	www.iobion.com
LabBook	eLabBook web-enabled electronic notebooks; annotated human genome database and data-mining tools	McClean, Virginia	www.labbook.com
LabVelocity	Jellyfish desktop bioinformatics software; information services	San Francisco, California	www.labvelocity.com
LION Bioscience	Bioinformatics software, database development and management; DiscoveryCenter platform for data integration; contract bioinformatics	Heidelberg, Germany	www.lionbioscience.com
MiraiBio	DNASIS desktop software for DNA sequence assembly and analysis, protein-sequence analysis, and analysis of microarray data	Alameda, California	www.mirai.bio.com
Molecular Biology Insights	Oligonucleotide identification software	Cascade, Colorado	www.oligo.net
Oracle	Database and data management solutions	Redwood Shores, California	www.oracle.com/industries/life_sciences/index.html?content.html
Paracel	Software for sequence assembly, analysis and sequence-based genotyping	Pasadena, California	www.paracel.com

Company	Products/activity	Location	URL
Premier Biosoft	Desktop bioinformatics packages for sequence analysis, primer design and two-hybrid protein interactions	Palo Alto, California	www.premierbiosoft.com
Redasoft	Visual Cloning 3 software for sequence management and analysis	Toronto, Ontario	www.redasoft.com
Rosetta BioSoftware	Rosetta Resolver and Rosetta Luminator gene-expression data-analysis system	Kirkland, Washington	www.rosettatabio.com
Sagittus Solutions	MaxD data warehouse, analysis and visualization software for microarray data	Manchester, UK	www.sagittussolutions.co.uk
Silicon Genetics	MetaMine, GeNet and GeneSpring microarray analysis software	Redwood City, California	www.sigenetics.com
Softberry	Software for sequence and genome analysis, and database searching	Mount Kisco, New York	www.softberry.com
Spotfire	DecisionSite analytical and statistical data-management software	Somerville, Massachusetts	www.spotfire.com
SPSS	Clementine statistical and data-mining software; LexiQuest literature text-mining and categorization tools	Chicago, Illinois	www.spss.com
Textco	Gene Construction Kit and Gene Inspector desktop bioinformatics packages and electronic lab notebook	West Lebanon, New Hampshire	www.textco.com

Computer systems, middleware and laboratory information management systems (LIMS)

Amersham Biosciences	Sierra Laboratory Workflow Systems for microarray, genotyping, sequencing and 2D-mass spectrometry data	Piscataway, New Jersey	www.amershambiosciences.com
CLONDIAG	PARTISAN microarray LIMS	Jena, Germany	www.clondia.com
geneticXchange	discoveryHub middleware platform for biological data integration	Menlo Park, California	www.geneticxchange.com
HeliXense	Genomics Research Network Architecture (gRNA) software and system-infrastructure supporting biological data management	Singapore	www.helixense.com
Hewlett-Packard	Computer systems, networks and data-management systems for the life sciences	Palo Alto, California	www.hp.com/techservers/index.html
InforSense	Kensington Discovery Edition research-management software platform; KDE TextSense	London, UK	www.inforsense.com
InnaPhase	LIMS for drug discovery	Philadelphia, Pennsylvania	www.innaphase.com
Labtronics	Laboratory software and control systems, LIMS interfacing	Guelph, Ontario, Canada	www.labtronics.com
L.I.M.S.	StarLIMS laboratory information management system	South Hollywood, Florida	www.starlims.com
Mitsui Knowledge Industry	LIMS; data management and analysis of gene-expression and SNP data	Tokyo, Japan	bio.mki.co.jp
NEC	Computer systems and networks	Tokyo, Japan	www.nec-global.com
NuGenesis	SDMS data-management system; scientific data management services	Westborough, Massachusetts	www.nugenesis.com
Predictive Patterns	GeneLinker gene-expression analysis software	Kingston, Ontario, Canada	www.predictivepatterns.com
ProteDyne	LIMS middleware for integration of network-enabled laboratory software	Martinsried, Germany	www.proteDyne.com
Silicon Graphics	SGL servers for high-throughput computing, visualization and data management	San Francisco, California	www.sgi.com
Sun Microsystems	Servers and workstations for high-throughput computing; universal software platforms for networks	Santa Clara, California	www.sun.com
TimeLogic	DeCypher system for accelerated bioinformatics	Crystal Bay, Nevada	www.timelogic.com
TurboWorx	Open computational platforms for biological research data including bioinformatics	Burlington, Massachusetts	www.turbogenomics.com

Services

Aber Genomic Computing	Design of data-mining and predictive modelling software	Aberystwyth, UK	www.abergc.com
AGOWA	Genome analysis; automated sequence annotation bioinformatics services	Berlin, Germany	www.agowa.de
Bioinformatics Services	Computational biology; bioinformatics services	Rockville, Maryland	www.bioinformatics-services.com
Chemical Computing Group	Bioinformatics software, services and computer-aided molecular design	Montreal, Quebec, Canada	www.chemcomp.com
Cyberell	Bioinformatics software and services	Helsinki, Finland	www.cyberell.com
ePitope Informatics	Epitope prediction over the web	Durham, UK	www.epitope-informatics.com
GeneData	Bioinformatics systems and services; database development and management	Basel, Switzerland	www.genedata.com
Keygene	DNA fingerprint analysis software; contract genomics and bioinformatics services	Wageningen, The Netherlands	www.keygene.com
SRI International	Contract informatics services	Menlo Park, California	www.sri.com
Tripos	Molecular modelling and virtual-screening software; contract informatics	St Louis, Missouri	www.tripos.com

General

Applied Maths	Gel fingerprint analysis and bioinformatics software; contract bioinformatics	Austin, Texas	www.applied-maths.com
Applied Biosystems	Instruments and assays for genomics, proteomics and molecular biology; myScience website for genome analysis	Foster City California	home.appliedbiosystems.com
Bioalma	Bioinformatics software, software development, consultancy and training	Madrid, Spain	www.almabioinfo.com
Bio-Rad	WorksBase bioinformatics software for proteomics	Hercules, California	www.discover.bio-rad.com
BioSolveIt	Molecular modelling, docking, and protein threading software; services	St Augustin, Germany	www.biosolveit.de
Dalicon	Bioinformatics software for large-scale data management and analysis	Nijmegen, The Netherlands	www.dalicon.com
Elsevier Science	Scirus web-search engine for scientists	Amsterdam, The Netherlands	www.scirus.com
Invitrogen	Kits and reagents for cloning, genomics, genotyping, proteomics, molecular and cell biology	Carlsbad, California	www.invitrogen.com
Partek	Pattern-recognition and interactive visualization software; consulting services	St Charles, Missouri	www.partek.com
Systat Software	Systat statistical and analytic graphics software; SigmaPlot and other graphing software	Richmond, California	www.systat.com

● see advertisement

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayouni (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Production Manager:

Billie Franklin
To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Stefan Hales

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan, Odo Wulffien

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffien@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Haraikatamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: rasami@naturejp.com

naturejobs

Balancing the books

A report evaluating how to attract more Europeans into scientific careers couldn't have come at a better time. Jose Mariano Gago, Portugal's former science minister, delivered the strategy document to European research commissioner Philippe Busquin this month — just weeks after leading senior French scientists voted to strike over the limited prospects for their younger colleagues (see *Nature* 428, 241; 2004).

The report and the strike both point to a missing ingredient in European research: money. Cuts to the French system and a need for reform triggered a call to arms from senior researchers expressing solidarity for their junior colleagues, who face relatively bleak prospects. And Gago's report served as a reminder that Europe won't become the leader in the knowledge economy that Busquin envisages without an accompanying increase in research investment to 3% of gross domestic product by 2010.

The report speaks of a shortage of scientists in Europe, particularly young ones. But it is likely that postgraduates are not aware of this shortfall — after all, the difference between facing 199 competitors for a job rather than 200 isn't much of a shift in odds. Gago's report calls for more awareness of posts outside academia as one way to address this problem. But without funds to support academia, young people won't be drawn to either conventional or alternative scientific careers — at least in Europe.

But the French protest could offer hope to all of Europe. Following the strikes, the opposition party made significant gains in the French regional election (see *Nature* 428, 454; 2004) — perhaps a message that the public has voted in favour of science. Ultimately, if the European Union wants to meet its research goals, the public will probably need to send similar messages in other European countries.

Paul Smaglik
Naturejobs editor



Contents

CAREER VIEW

Bricks & Mortar

Arizona plots a course

Graduate Journal

Losing control

Movers

Emilio Emini

p782

WWW.NATUREJOBS.COM

Career centre

Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

GRADUATE JOURNAL

Losing control

Now that I've settled on Paris as my postdoc destination, the hard part begins — the mad rush towards graduation. As I'm working feverishly in the lab and on my computer, trying to compile four years' data into a cohesive 'story', I pray that my thesis committee decides that I've done enough. But how much is enough?

Only my thesis committee can answer that question. I am completely at their mercy. Counting on them to guide me through my defence feels like falling backwards with my eyes closed and trusting the four committee members to catch me before I hit the ground. This abdication of control is a little hard to swallow, but I have no choice other than to go with the committee's collective judgement.

For some students, this does not come easily. I've heard horror stories about hard-line advisers black-balling students for years over seemingly frivolous details. As my decisive committee meeting approaches, I feel fortunate because I respect my committee members and have grown quite close to some of them. I'm grateful for their guidance, and I just hope they steer me through a smooth defence and let me move on by this summer. There's a chocolate croissant with my name on it somewhere in Paris and I'm anxious to take a bite. ■

Tshaka Cunningham is a fifth-year graduate student at Rockefeller University in New York.

BRICKS & MORTAR

Arizona plots a course

The next six years will see the University of Arizona in Tucson expand its life-sciences capabilities. As part of a broader initiative to boost bioscience research in the state, the university plans to take on some 100 new faculty members in medicine, pharmacology, public health, agriculture and engineering.

The recruitment strategy, submitted to the university's vice-president for research at the beginning of the month, follows hot on the heels of a ten-year 'road-map' for the state unveiled last month. This was assembled by a statewide consortium of leaders from Arizona's government, education and business sectors who want to see the state focus its research efforts on cancer, neuroscience and bioengineering. The consortium's plan, which encompasses both the

private sector and the state's universities, anticipates generating up to 32,000 bioscience jobs in Arizona by 2014.

A committee at the University of Arizona reacted to this vision with its own plan, which adds increases in research into diabetes, obesity and asthma to the state's core areas. Michael Cusanovich, the committee's chairman and director of the university's Arizona Research Laboratories, foresees hiring 30 members of staff in the first year of the plan.

Some of these new recruits will be in what Cusanovich calls the "crosscutting" disciplines of bioimaging and mathematical biology. These positions would support work being done in the three targeted research areas.

Although the committee is still discussing the improvements needed to house the new faculty

members, it is likely that some space currently under construction will come into play. The university broke ground on two life-science projects this year — the \$60-million Institute for Biomedical Science and Biotechnology, which will emphasize interdisciplinary research, and the \$54-million Medical Research Building, which will house translational research. Both are scheduled for completion in 2005 and there are tentative plans for a third life-sciences building on the campus.

The six-year plan does not yet have an estimated cost, and it is also unclear how many of the proposed new positions would be net gains, as opposed to a reallocation of existing faculty slots.

Nevertheless, the university's expansion would complement other efforts in bioengineering, neuroscience and cancer research across the state. ■

Paul Smaglik is editor of Naturejobs.

MOVERS Emilio Emini, senior vice-president, International AIDS Vaccine Initiative, New York



A broad interest in biology and economic necessity combined to steer Emilio Emini into graduate school. He had cultivated a general interest in the life sciences during his undergraduate years, but when he graduated, he wasn't sure what to do with it. With a recession under way, employment opportunities were scarce.

"What do you do when job prospects are bleak?" asks Emini. "You go to grad school." He won a National Science

Foundation pre-doctoral fellowship and headed for Cornell University. Passing through a number of labs during his first year, Emini ended up in one working on a strain of Venezuelan encephalitis virus. He found exploring the pathogenesis of different strains fascinating and became hooked on virology.

For his postdoc, Emini moved to the State University of New York at Stony Brook, primarily because Eckard Wimmer, an expert on poliovirus worked there. At the time, polio was one of the most understood viruses, so it made a good starting point to explore virology further.

When Emini arrived, Wimmer was finishing the genome sequence of poliovirus. That, coupled with the advent of monoclonal antibodies, meant that Emini could explore biological questions with brand new tools — a chance that had him feeling "like a kid in a candy shop".

Emini switched to industry in 1983 when Merck's expansion in vaccine

research matched his rising interest in the field. He didn't expect to stay in industry, but "the science kept getting interesting", he says. He worked on HIV as the company moved into AIDS research during the mid-1980s and helped to develop the multidrug cocktails that now allow many people to live with the virus.

He rose through the ranks at Merck, but found that as a senior vice-president of vaccine research, he could no longer focus on HIV. And, as a manager, his schedule often kept him from the lab.

As he approached his fiftieth birthday, he realized that he wanted to focus on HIV, and that he needed 10–15 years to make an impact. "It was an alignment of the planets," Emini says.

Taking over as senior vice-president at the International AIDS Vaccine Institute in New York, he is in some ways back where he started — looking at a big problem and hoping he can use new technologies to crack it. ■

CV **1983–2004:** Merck Research Laboratories, West Point, Pennsylvania (starting as senior research microbiologist, rising to senior vice-president of vaccine and biologics research in 2002)

1980–83: Postdoctoral fellow, Department of Microbiology, State University of New York at Stony Brook

1980: PhD in microbiology, Cornell University, Ithaca, New York